Psychometrika

CONTENTS

(cont.)

VOLUME TWENTY-FIVE SEPTEMBER 1960 NUMBER 3

Psychometrika

CONTENTS (*Cont.*)

VOLUME TWENTY-FIVE SEPTEMBER 1960 NUMBER 3

PSYCHOMETRIKA—VOL. 25, NO. 3
SEPTEMBER, 1960

ON A CONNECTION BETWEEN FACTOR ANALYSIS AND MULTIDIMENSIONAL UNFOLDING*

CLYDE H. COOMBS
UNIVERSITY OF MICHIGAN

AND

RICHARD C. KAO
PLANNING RESEARCH CORPORATION
LOS ANGELES, CALIFORNIA

Given the preference ordering of each of a number of individuals over a set of stimuli, it is proposed that if the preference orderings are generated in a Euclidean space of r dimensions which can be recovered by unfolding the preference orderings, then a factor analysis of the correlations between individual's preference orderings will yield a space of $r + 1$ dimensions with the original r-space embedded in it, and the additional dimension will be one of social utility. The proposition is clearly shown to be satisfied by means of the Monte Carlo technique for both random and lattice stimuli in three dimensions and for two other examples with random stimuli in one and two dimensions.

The unfolding technique for preferential choice behavior [7, 8] is derived from a model of the following form. An individual, in making preferential choices among a set of alternatives, may be represented by a point in an r-dimensional Euclidean space, E^r, and correspondingly, each alternative may be represented by a point in the same space. The individual prefers one alternative to another if and only if the point corresponding to the preferred alternative is nearer to the point corresponding to the individual. To each point corresponds an r-tuple which is a set of measures on the dimensions spanning the space. These dimensions may be interpreted as psychological variables generating the preferences of the individuals, where the point corresponding to an individual is an ideal point representing a hypothetical alternative preferred to all other possible ones. Inconsistency

*The preparation of this paper was supported in part by a grant from the National Science Foundation and in part by Project MICHIGAN, a project of the University of Michigan in the field of Combat Surveillance sponsored by the Department of the Army. The contract (DA-36-039 ac 78801) is administered by the U. S. Army Signal Corps. The authors are indebted to L. A. Raphael, Caroline K. Tefft, and F. M. Goode for programming assistance, and to L. W. Staugas for providing other computer services during various stages of this study.

of preferences, to be distinguished from intransitivity, may be generated by random variability in the locus of points [10].

According to the model, an individual's dominant preferences may be represented by a rank order scale of the alternatives given by the transitive set of stochastically determined pairwise preferences. Such a scale is called an I scale and may be regarded as folding the space by picking it up at the ideal point and collapsing it into a line with the measure of the stimulus points on this line corresponding to their respective distances from the ideal point. Distinct ideal points generate distinct I scales in this manner. With ordinal preference data such I scales have a many-one mapping into equivalence classes corresponding to distinct rank order I scales. The unfolding technique is the name given to the method for determining the number of dimensions and the rank order of the projections on the dimensions, and, in the case of one dimension, ordered metric information.

The problem of determining the dimensionality of a Euclidean space in which a set of I scales may be unfolded was solved by Bennett [6] and the problem of determining the configuration of the points for both stimuli and individuals (called a Joint space) was solved by Hays [11].

The following problem naturally arises. Suppose one intercorrelated the individuals' I scales and factor analyzed; what relation would the factorial solution have to the E^r assumed to have generated the preferences?

The Proposition

Consider the simple case of a one-dimensional latent attribute generating the preferences of individuals over a set of alternatives. The ideal points of the individuals and the points for the alternatives are all points on a line, a Joint scale. To avoid sampling fluctuations, assume the stimulus points are dense and that the two sets of points range over the same segment of the line.

Consider the I scale of an individual (A) at the extreme left end of the scale and that of another individual very close to him. Clearly, their preference orderings will be almost identical and will correlate close to plus one. The I scale for individual A will correlate progressively less with I scales of other individuals as they are farther removed from him on the Joint scale. In fact, the correlation will be zero between individual A and the median individual in the distribution, and will ultimately be minus one between him and the individual at the extreme opposite end of the scale. The median individual will have correlations ranging from close to plus one with those individuals near him on either side, to zero with the individuals at either end.

Clearly, if each individual is represented by a unit vector from a common origin and the correlation between individuals by the cosine of the angle between the corresponding vectors, the configuration corresponding to the correlation matrix will be a semicircle with the individuals corresponding to a fan of vectors such that the vector of the median individual projects verti-

cally upward and orthogonal to the vectors of the two extreme individuals which form an angle of 180 degrees. The order of the termini of the vectors on the arc would correspond exactly to the order of the corresponding points on the original line.

If one factor analyzes such a configuration by the method of principal components, the first dimension would be the original line which generated the preferential choices; the *second* dimension would be the vector of the median individual on the line. On the latter dimension the projections of individual points are in reverse order with respect to how closely each is to all the others on the line. In another context, this second dimension is called a social utility [9]. The higher the projection, the more nearly that point best represents all the other points in the sense of being nearest to them all.

If we consider the case of a two-dimensional latent attribute space generating the preferential choices, we now have two superimposed bivariate distributions—one for individuals and one for stimuli. If one considers the correlation of the I scale of an individual on the rim of this space with other individuals, it seems reasonable that the correlations will progressively decrease through zero to minus one as one approaches an individual across the space from him, and that the median individual on the plane will correlate non-negatively with everyone. The configuration generated by the set of unit vectors is now a hemisphere in three dimensions, with the median individual represented by a unit vector perpendicular to the plane in which the vectors of all individuals on the rim of the plane lie. If such were the case, a factor analysis would yield three dimensions, with the third principal component again corresponding to a social utility and the first two dimensions representing the original space which generated the preferential choices.

While not as intuitively obvious, we may generalize this proposition to a space of r dimensions in which we would expect the configuration corresponding to the correlation matrix to be a semihypersphere in $r + 1$ dimensions; the $(r + 1)$th principal component would be a social utility and the first r dimensions would correspond to the original space.

This proposition was first conjectured by the first author but later more fully studied by the second using the Monte Carlo technique. Any attempt to realize the idealized version of the proposition would necessarily lead to some distortion, the matching of the two being sensitive to the density of stimulus points and the joint distribution of stimulus and individual points and to the measure used for the correlation between two individuals' I scales. In practice only a finite number of stimulus points can be used so the working definition of a genotypic space is the chosen finite set of stimulus points. The theorem which is conjectured is this: given m arbitrary points in E^n, then they lie in an r-subspace of E^n if and only if with probability 1, the rank of the product moment correlation matrix approaches $r + 1$ as the number of stimulus points approaches infinity.

Imbedding of Genotypic Space into Factor Space

In order to test the plausibility of the proposition discussed above under rather general and varying conditions, several problems were constructed and explored, of which two related ones in three dimensions play the major role. These will be presented first.

Three sets of 15 random numbers are taken to represent the coordinates of 15 individuals in E^3, and another three sets of 30 random numbers, those of 30 stimuli in the same space (Tables 1 and 2). (All numbers [14] were

TABLE 1

Coordinates of Individual Points in E^3

	a	b	c
01	-0.47883	-0.12812	0.30109
02	-0.20438	-0.40540	0.26483
03	-0.49558	0.23608	-0.18766
04	0.41039	-0.42816	-0.29035
05	-0.45173	0.54625	0.41746
06	0.08085	0.29372	-0.04339
07	-0.26920	-0.34540	0.07160
08	0.27289	0.32257	-0.36360
09	0.05593	-0.13210	-0.33086
10	0.15816	0.00408	-0.34882
11	0.40540	-0.27578	-0.23506
12	-0.46175	-0.39914	0.01397
13	0.25472	0.54289	-0.32123
14	0.30705	-0.05145	0.48096
15	0.41614	0.22003	0.49106

first taken to be seven-place decimal fractions and computations carried out in this manner, but rounded to five places after the completion of the study.) Since all numbers were decimal fractions, the Joint space for both individuals and stimuli is, by definition, a cube in E^3 with length of its sides equal to 2 and center at the origin, called the basic *cube*. A third set of points is taken to represent a second set of stimulus points, these being the 64 lattice points of a "grid" contained in the basic cube. On each dimension the points take on one of the four values $-.6$, $-.2$, $+.2$, $+.6$, yielding $4^3 = 64$ points. For simplicity, we shall distinguish the two sets of stimulus points by calling them *random stimuli* and *lattice stimuli*, respectively. The motivation for taking the latter is twofold: (i) to see if an increase in the number of stimuli used would yield a better fit to the idealized situation, and (ii) to test if the model were feasible with quite arbitrary selection of stimulus points, random as well as nonrandom.

TABLE 2

Coordinates of Random Stimulus Points in E^3

	a	b	c
01	0.03991	-0.40188	0.28193
02	-0.38555	-0.34414	0.32886
03	0.17546	0.10461	0.39510
04	-0.32643	-0.52861	0.27699
05	-0.24122	-0.30231	-0.10274
06	0.30532	0.21704	-0.35075
07	-0.03788	0.42402	0.56623
08	0.48228	-0.07405	-0.36409
09	-0.32960	0.53845	0.57620
10	-0.19322	-0.57260	0.07399
11	-0.11220	-0.47744	-0.14454
12	0.31751	-0.48893	0.07481
13	-0.30934	0.16993	0.27499
14	0.22888	0.33049	-0.35902
15	-0.41849	-0.08337	-0.46850
16	-0.46352	0.36898	0.14013
17	-0.11087	-0.48297	0.56303
18	-0.52701	-0.19019	0.39904
19	0.57275	0.32486	0.45134
20	-0.20857	0.01889	0.37239
21	0.15633	0.07629	-0.18637
22	-0.38688	0.43625	-0.05327
23	0.25163	-0.11692	0.43253
24	0.36815	0.25624	-0.53342
25	-0.04515	0.06345	-0.13574
26	0.14387	-0.00008	-0.29593
27	0.51321	0.55306	-0.44989
28	0.05466	0.18711	0.52162
29	-0.39528	-0.16120	0.04737
30	-0.07586	-0.04235	0.16894

The Euclidean distances of each individual from all the stimuli (random or lattice) are computed and these measures provide an I scale for the individual, which is a ratio scale rather than an ordinal scale. The product moment correlations are then computed between each pair of individuals' I scales yielding two correlation matrices, one for the random stimuli M_r, (Table 3) and one for the lattice stimuli M_l (Table 4). These correlation matrices, with unity in the diagonal are then factored by the method of principal components. Two different subroutines (IBM 704 and RAND JOHNNIAC) were used independently to duplicate all computations. The

TABLE 3

Correlation Between I Scales for Random Stimuli

	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15
01	1.00000														
02	0.81806	1.00000													
03	0.40572	0.10542	1.00000												
04	-0.36616	0.11950	-0.20383	1.00000											
05	0.53181	0.02277	0.52189	-0.78366	1.00000										
06	-0.33222	-0.46428	0.40337	0.20512	0.18277	1.00000									
07	0.76629	0.94362	0.32694	0.22961	-0.01799	-0.32655	1.00000								
08	-0.73815	-0.67589	0.08695	0.46280	-0.33930	0.78433	-0.51320	1.00000							
09	-0.26828	0.02011	0.30207	0.83379	-0.50498	0.51645	0.25743	0.63858	1.00000						
10	-0.51575	-0.29705	0.19385	0.76176	-0.48119	0.69258	-0.08422	0.86011	0.93347	1.00000					
11	-0.46329	-0.02397	-0.19668	0.98148	-0.73770	0.36014	0.08807	0.59572	0.84772	0.83680	1.00000				
12	0.78772	0.93365	0.41895	0.13480	0.03176	-0.36574	0.98322	-0.53188	0.22048	-0.12251	-0.01156	1.00000			
13	-0.74154	-0.79191	0.11925	0.25679	-0.16748	0.81675	-0.64825	0.96836	0.46858	0.73384	0.41059	-0.65380	1.00000		
14	0.26843	0.34916	-0.37285	0.02601	0.17020	0.05328	0.13859	-0.30203	-0.23865	-0.25308	0.04607	-0.00194	-0.29391	1.00000	
15	0.05611	-0.12378	-0.38483	-0.14621	0.27643	0.30513	-0.31281	-0.02364	-0.32576	-0.18135	-0.04866	-0.43069	0.06526	0.85786	1.00000

numbers from these two sources agree to seven decimal places except for signs, i.e., a characteristic vector from one subroutine may be the reflection of another from the other subroutine.

The characteristic values, λ_i, for the two correlation matrices are given in Table 5. It can be seen that a sharp drop in the magnitude of the characteristic value occurs after the fourth one. We take, therefore, the first four columns of the factor matrices (Tables 6 and 7) as factor loadings or co-

TABLE 4

Correlation Between I Scales for Lattice Stimuli

	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15
01	1.00000														
02	0.80180	1.00000													
03	0.45996	0.09669	1.00000												
04	-0.34670	0.18399	-0.35755	1.00000											
05	0.61419	0.09619	0.59587	-0.70245	1.00000										
06	-0.00221	-0.16857	0.44345	0.05399	0.45912	1.00000									
07	0.76486	0.93343	0.34352	0.23930	0.07586	-0.08260	1.00000								
08	-0.50001	-0.48321	0.20553	0.36304	-0.06191	0.80035	-0.32676	1.00000							
09	-0.13477	0.16287	0.28975	0.75084	-0.35724	0.40084	0.39914	0.60457	1.00000						
10	-0.33440	-0.10284	0.22977	0.70250	-0.30290	0.60162	0.10993	0.83923	0.93606	1.00000					
11	-0.37094	0.12037	-0.32828	0.97985	-0.61672	0.22331	0.16919	0.50019	0.76455	0.77163	1.00000				
12	0.76137	0.84042	0.46245	0.09010	0.09797	-0.18764	0.95990	-0.39451	0.33566	0.02720	-0.00270	1.00000			
13	-0.45148	-0.58134	0.27127	0.09957	0.14840	0.84892	-0.45322	0.95522	0.38240	0.65884	0.22527	-0.49730	1.00000		
14	0.26537	0.42697	-0.38849	0.20445	0.17099	0.28854	0.17630	-0.03695	-0.10708	-0.05619	0.28848	-0.06571	-0.04201	1.00000	
15	0.05755	0.06988	-0.33573	0.04294	0.29987	0.49538	-0.15479	0.19408	-0.19580	-0.02219	0.17765	-0.37016	0.26100	0.90801	1.00000

TABLE 5

Characteristic Values for M_r and M_ℓ

	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15
λ_i, Random Stimuli	6.39715	4.01020	2.35246	1.99655	0.39150	0.10905	0.06822	0.03352	0.02680	0.01196	0.00999	0.00183	0.00075	0.00034	
λ_i, Lattice Stimuli	5.21363	3.94788	3.44330	2.49448	0.12577	0.07310	0.03068	0.02790	0.02052	0.01096	0.00515	0.00385	0.00161	0.00104	0.00014

ordinates of the 15 individuals in four dimensions. The statistical theory for testing the number of significantly positive characteristic roots of a sample correlation matrix has yet to be worked out (cf. [1], p. 330). Our investigation lends some convincing evidence that such a theory can be developed [cf. 1, 2, 3, 4, 5, 12].

Two crucial questions arise. First, how are the original coordinates of

TABLE 6

Principal Components Factor Loadings for Random Stimuli

	01	02	03	04
	6.39715	4.01020	2.35246	1.99655
01	0.85262	0.27021	-0.22809	0.33375
02	0.69980	0.62881	0.17248	0.26119
03	0.08652	0.25101	-0.92805	0.20117
04	-0.57057	0.70975	0.38252	0.06207
05	0.47920	-0.48049	-0.56888	0.39698
06	-0.68648	-0.08115	-0.41818	0.54391
07	0.55962	0.78716	-0.04066	0.21914
08	-0.96190	0.00826	-0.20677	0.03215
09	-0.63852	0.73509	-0.16018	0.13697
10	-0.85590	0.46673	-0.15916	0.12525
11	-0.69323	0.60280	0.34346	0.12922
12	0.59545	0.76164	-0.16490	0.11234
13	-0.92659	-0.28631	-0.18912	0.26799
14	0.25163	-0.14168	0.57725	0.74380
15	-0.10412	-0.56050	0.39376	0.76243

TABLE 7

Principal Components Factor Loadings for Lattice Stimuli

	01	02	03	04
	5.21363	3.94788	3.44330	2.49448
01	0.72719	0.46805	-0.45787	0.14273
02	0.53410	0.78105	-0.01001	0.29306
03	0.10687	0.26099	-0.76577	-0.54484
04	-0.59226	0.60106	0.48881	0.15722
05	0.37333	-0.17152	-0.87300	0.12318
06	-0.58528	0.09476	-0.77412	0.19804
07	0.41858	0.89902	-0.09576	0.00533
08	-0.92210	0.02215	-0.35550	-0.07667
09	-0.61730	0.73370	-0.02764	-0.25316
10	-0.83504	0.50160	-0.12971	-0.16607
11	-0.69580	0.55692	0.36150	0.23753
12	0.51236	0.81177	-0.10147	-0.22775
13	-0.81839	-0.18864	-0.51428	-0.05858
14	0.00093	0.18049	-0.10485	0.97188
15	-0.18760	-0.12257	-0.27787	0.92650

the individual points in three dimensions related to their factor loadings in four dimensions? Second, what is the significance of the "extra" dimension obtained?

According to the proposition, the configuration corresponding to the correlation matrix is a set of unit vectors in E^4 whose projections in a subspace E^3 orthogonal to the median individual will faithfully reproduce the configuration of the individual points in the original genotypic space. Hence, the first question can be settled if we show that Table 1 can be "imbedded" into Table 6 and into Table 7.

To this end, Tucker's method of congruence is used [15]. His coefficient of congruence, Q_r , is similar to a product moment correlation between the loadings on factor r in the factor space and those in the original space.The values of Q_r for each of the three original dimensions as recovered by the two factor analyses are given in Table 8. The congruence appears reasonably

TABLE 8

Congruence of Original Dimensions with Factor Space

	r_1	r_2	r_3
Q_r, Random Stimuli	.99492	.97699	.98790
Q_r, Lattice Stimuli	.99154	.98254	.99687

high and a good fit of the original configuration of individual points into a three-dimensional subspace of the factor space is possible.

The Extra Dimension in the Factor Space

According to the proposition, the genotypic space can be imbedded in the factor space; the factor space will have an additional dimension and the projection of a point on this extra dimension will be related to how close each point was to all the other points in the genotypic space. The first two parts of the proposition have been sustained by the results reported above and it remains now to test the last part.

The projection of each vector on the extra dimension of the factor space is readily given knowing the length of the vector in the factor space of four dimensions and its reduced length in the three-dimensional subspace that corresponds to the original genotypic space.

The average distance of any point from all the others in the original genotypic space is readily obtained from Table 1. The smaller the average distance of a point from all the others the nearer the point lies to the median of the population and hence the higher its projection on the extra dimension

of the factor space. The Spearman rank order correlations between average distances in genotypic space (ordered from smallest to largest) with projections on the extra dimension (ordered from largest to smallest) is .723 and .896 for random and lattice stimuli respectively, significant at the .005 level. It follows, therefore, that there is reasonable evidence for answering in the affirmative both questions which led us to include a second set of stimulus points in the three-dimensional problem.

Two more problems were run to test the proposition when the genotypic space is of dimension 1 or 2. For this purpose, only one set of stimulus points was retained by pairing off the first column in Table 1 against that in Table 2, or the first two columns in Table 1 against those in Table 2. Euclidean distances between individuals and random stimuli in 1 and 2 dimensions were first computed and then the correlation matrices of individuals over stimuli, which were factored by the method of principal components. Only a summary of the results are presented here. The first five (and largest) characteristic values for the one-dimensional case and the two-dimensional case are presented in Table 9.

TABLE 9

Characteristic Values for the One- and Two-Dimensional Genotypic Space

	1	2	3	4	5
One-Dimensional	12.02233	2.70781	0.13125	0.10578	0.01797
Two-Dimensional	7.38998	4.81759	2.36532	0.24135	0.07356

A sharp drop in the magnitude of the characteristic value occurs after the second for the one-dimensional case and after the third for the two-dimensional case, indicating that the factor space for preferences had one additional dimension beyond the genotypic space which generated the preferences. Again Tucker's method is used for maximal congruence and the Q_r for the one-dimensional case is 0.976 and for the two-dimensional case are 0.989 and 0.986 for the first and second dimensions respectively. Spearman rank order correlations between average distance of an individual's point from all the others in the genotypic space and the projection of the individual on the extra dimension were 0.761 and 0.669, significant at the 0.005 level, for the one- and two-dimensional cases respectively.

Discussion

We recapitulate briefly the main results of the preceding two sections. A Joint space is taken with both individuals and stimuli as points in it. An

I scale of preferences over the stimuli is constructed for each individual by taking the Euclidean distances of these stimuli from the individual's ideal point. Correlating individuals' *I* scales gives rise to a matrix of correlations which are factored by the method of principal components. In each problem, the dimension of the factor space is noticeably one higher than the original genotypic space. But, the configuration of the individual points in the original genotypic space can be faithfully reproduced in a hyperplane of the factor space. The rank orders of the projections of the individual vectors on the extra dimension correlate highly in reverse order with those of their genotypic space. These results are obtained when the Joint space is of different dimensions and the stimulus points are quite arbitrarily chosen.

There are several aspects which need to be discussed because of their relevance to the practical application of the propositions tested here. In any practical application there would be two sources of error or distortion, one of which is present in this study. The first is that the basic data would normally consist of rank order preference scales rather than the actual distances to stimulus points. This means that the product moment correlation can only be approximated. The second is that the distribution of stimulus points relative to that for the individuals can distort the factor space. This is most obviously evident in the one-dimensional case in which the stimuli that lie between two individuals tend to produce negative correlation between their preference orderings and the stimuli that lie outside of them tend to produce positive correlation. Clearly if the density of the stimulus points between two individuals is unusually high or low, the correlation between their preferences will be biased toward negative or positive correlation, and they will appear in the factor space as farther apart or nearer together than in the genotypic space.

A further aspect relevant to practical application is that in the real case one arrives first at the factor space and seeks the genotypic space. This requires determining the extra dimension in the factor space, with no prior knowledge of the genotypic space, and then rotating it out in order to work with just the genotypic space that remains. The following argument suggests how this may be done.

Our model states that all individual vectors in the genotypic space are "blown up" into unit vectors whose termini lie on a semihypersphere bounded by a hyperplane containing the genotypic space. In this process, the distance in the genotypic space of an individual from the median individual is changed by a monotone transformation into the distance on the semihypersphere between the termini of the unit vectors representing these individuals in the factor space. Therefore, the rank orders of the distances of all individuals from the median individual will not be affected. If a rotation about the origin is made of all individual vectors in the factor space, these rank orders still remain invariant. This means that we may determine the

social utility dimension in the factor space in exactly the same manner as we do in the genotypic space. That is, we use the coordinates of individuals in the factor space and find the median individual accordingly. The social utility dimension is then passed through this individual and the projections of other individuals on this dimension can be computed. By our observation above, the rank orders of all individuals from this median individual should correlate highly with those in the original genotypic space.

Summary and Conclusions

A model called the unfolding technique for analyzing preferential choice data assumes that individuals and stimuli may be represented by points in a Euclidean space of r dimensions and that an individual's preference ordering of the stimuli reflects the order of their increasing distance from his position in the space. Such a preference ordering is called an I scale. Given the I scales of a number of individuals, methods are available for determining the dimensionality of the space and the configuration of points in the space.

On the other hand, correlations between the preference orderings of individuals could be computed and the resulting correlation matrix factor analyzed. Naturally arising then is the question of what the relation would be between the genotypic space which gives rise to the preference orderings, and is recovered by the unfolding technique, and the space obtained by multidimensional factor analysis.

A heuristic argument was presented for the following propositions:

(i) if the genotypic space is Euclidean with r dimensions, the factor space will have $r + 1$ dimensions;

(ii) the genotypic space can be imbedded in the factor space;

(iii) the additional dimension in the factor space will be a social utility dimension in the sense that the nearer a point is to all the other points in the genotypic space the higher its projection is on this extra dimension in the factor space.

The problem was studied by the Monte Carlo technique. Three sets of 15 random numbers were taken as the coordinates of 15 individuals in E^3 and three sets of 30 random numbers, those of 30 stimuli in the same space. A second set of stimuli points was taken as the 64 lattice points of a cube 2 units on a side with center at the origin. Given this genotypic space, preference scales of individuals were computed for the random and for the lattice stimuli, correlation matrices between individual's preferences were obtained and factored by the method of principal components. This procedure was carried out for both sets of stimuli with $r = 3$ and with only the random stimuli with $r = 1$ and $r = 2$.

Tucker's method was used to test for congruence of the genotypic and factorial spaces. All three propositions were confirmed for both random and lattice stimuli with some slight superiority in favor of the lattice stimuli.

This could be due to the larger number of lattice stimuli or the regularity of their distribution or both.

The social utility dimension in the factor space was discussed including a possible method for isolating it.

The most general practical consequence of this development is that the methods of multiple factor analysis are revealed to be suitable for the discovery of the latent attribute variables underlying preferences after the social utility dimension has been removed, with the qualification that there will be some sensitivity to the density and the distribution of stimulus points in the space. A recent study by MacRae [13] is a case in point and the theory and technique developed here would have been useful in that study.

REFERENCES

[1] Anderson, T. W. *An introduction to multivariate statistical analysis.* New York: Wiley, 1958. Ch. V.

[2] Bartlett, M. S. Tests of significance in factor analysis. *Brit. J. Psychol., Statist. Sec.,* 1950, **3**, 77–85.

[3] Bartlett, M. S. The effect of standardization on a χ^2-approximation in factor analysis. *Biometrika,* 1951, **38**, 337–344.

[4] Bartlett, M. S. A further note on tests of significance in factor analysis. *Brit. J. Psychol., Statist. Sec.,* 1951, **4**, 1–2.

[5] Bartlett, M. S. Factor analysis in psychology as a statistician sees it. *Uppsala symposium on psychological factor analysis.* Uppsala, Sweden: Almqvist and Wiksell, 1953. Pp. 23–34.

[6] Bennett, J. F. and Hays, W. L. Multidimensional unfolding: determining the dimensionality of ranked preference data. *Psychometrika,* 1960, **25**, 27–44.

[7] Coombs, C. H. Psychological scaling without a unit of measurement. *Psychol. Rev.,* 1950, **57**, 145–158.

[8] Coombs, C. H. *A theory of psychological scaling.* Engin. Res. Inst. Bull., No. 34. Ann Arbor: Univ. Michigan Press, 1952.

[9] Coombs, C. H. Social choice and strength of preferences. In R. M. Thrall, C. H. Coombs, and R. L. Davis (Eds.), *Decision processes.* New York: Wiley, 1954. Pp. 69–86.

[10] Coombs, C. H. Inconsistency of preferences in psychological measurement. *J. exp. Psychol.,* 1958, **55**, 1–7.

[11] Hays, W. L. and Bennett, J. F. Multidimensional unfolding: determining configuration from complete rank order preference data. *Psychometrika.* (in press)

[12] Kendall, M. G. and Smith, B. B. Factor analysis. *J. roy statist. Soc.* (*B*), 1950, **12**, 60–94.

[13] MacRae, D., Jr. A factorial analysis of political preferences. *Revue Francaise de Science Politique,* 1958, **8**, 95–109.

[14] RAND Corporation. *A million random digits with* 100,000 *normal deviates.* Glencoe, Illinois: Free Press, 1955.

[15] Tucker, L. R. *A method for synthesis of factor analysis studies.* Princeton: Educ. Test. Serv. Res. Bull. No. 984, 1951.

Manuscript received 3/25/59

Revised manuscript received 11/19/59

PSYCHOMETRIKA—VOL. 25, NO. 3
SEPTEMBER, 1960

SOME ASYMPTOTIC PROPERTIES OF LUCE'S BETA LEARNING MODEL*

JOHN LAMPERTI AND PATRICK SUPPES

APPLIED MATHEMATICS AND STATISTICS LABORATORIES
STANFORD UNIVERSITY

This paper studies asymptotic properties of Luce's beta model. Asymptotic results are given for the two-operator and four-operator cases of contingent and noncontingent reinforcement.

For application to various simple learning situations, Luce and his collaborators, Bush and Galanter, [1, 7] have considered a learning model in which the changes in probability of response from trial to trial are not linear functions of the probability of response on the preceding trial. Both theoretical and empirical considerations have motivated the development of the beta model. Some learning theorists like Hull and Spence believe that overt response behavior may best be explained in terms of a construct like that of response strength. From this viewpoint stochastic learning models which postulate a linear transformation of the probability of response from one trial to the next, with the transformation depending on the reinforcing event, are unsatisfactory in so far as they offer no more general psychological justification of their postulates. From an empirical standpoint there is evidence in some experiments, particularly certain T-maze experiments with rats, that the linear stochastic models do not yield good predictions of actual behavior [1, 7].

On the basis of some very simple postulates [7] on choice behavior, Luce has shown that there exists a ratio scale v over the set of responses with the property that

$$p_{i,n} = \frac{v_n(i)}{\sum_i v_n(i)},$$

where $p_{i,n}$ is the probability of response A_i on trial n, and $v_n(i)$ is the strength of this response on trial n. Additional simple postulates lead to the result that the $v_n(i)$ are transformed linearly from trial to trial, and this unobservable stochastic process on response strengths then determines a stochastic process

*This research was supported in part by the Group Psychology Branch of the Office of Naval Research and in part by the Rockefeller Foundation.

in the response probabilities. Superficially, it would seem that the simplest way to study the asymptotic behavior of the response probabilities—a subject of interest in connection with nearly any learning data—would be to determine the asymptotic behavior of the response strengths $v_n(i)$ and then infer by means of the equation given above the behavior of the response probabilities. This course is pursued rather far by Luce [7] and encounters numerous mathematical difficulties. We have taken the alternative path of studying directly the properties of the nonlinear transformations on the response probabilities to obtain results on their asymptotic behavior.

We restrict ourselves to situations in which one of two responses, A_1 and A_2, is made. Let p_n be the probability of response A_1 on trial n, and let E_1 be the event of reinforcing response A_1, and E_2 the event of reinforcing response A_2.

Luce's beta model is then characterized by the following transformations: if A_j and E_k occurred on trial n, then for $j = 1, 2$ and $k = 1, 2$,

$$p_{n+1} = \frac{p_n}{p_n + \beta_{jk}(1 - p_n)}, \tag{1}$$

where $\beta_{jk} > 0$. Luce [7] gives a more general formulation. (Generally, we want $\beta_{j1} < 1$ and $\beta_{j2} > 1$, to reflect the primary effects of reinforcement; moreover, it is ordinarily assumed that $\beta_{11} < \beta_{21} < \beta_{12} < \beta_{22}$.) Throughout this paper it is assumed that $0 \neq p_1 \neq 1$.

The most important fact about (1) is that the operators commute. For example, suppose in the first n trials there are b_1 occurrences of A_1E_1, b_2 occurrences of A_2E_1, b_3 occurrences of A_1E_2, b_4 occurrences of A_2E_2; then it is easily shown that

$$p_{n+1} = \frac{p_1}{p_1 + \beta_{11}^{b_1}\beta_{21}^{b_2}\beta_{12}^{b_3}\beta_{22}^{b_4}(1 - p_1)}. \tag{2}$$

The aim of the present paper is to study asymptotic properties of the beta model for certain standard probabilistic schedules of reinforcement. The methods of attack used by Karlin [4] and by Lamperti and Suppes [6] for linear learning models do not directly apply to the nonlinear beta model.

The basis of our approach is to change the state space (the probability p_n is the state) from the unit interval to the whole real line in such a way that the transformations (1) become simply translations. The noncontingent case (the next section) then reduces to sums of independent random variables; the contingent cases can also be studied by "comparing" the resulting random walks with the case of sums of random variables. The probabilistic tool for this is developed and applied in later sections. The general conclusion to be drawn from our results is that for all but one case of noncontingent reinforcement individual response probabilities are ultimately either zero or one, which is in marked contrast to corresponding results for linear learning

models. Absorption at zero or one also occurs for many, but not all, cases of contingent reinforcement.

Noncontingent Reinforcement with Two Operators

If the probability of a reinforcement is independent of response and trial number, we have what is called simple noncontingent reinforcement. Let π be the probability of an E_1 reinforcement, and for simplicity let

$$\begin{cases} \beta_{11} = \beta_{21} = \beta, \\ \beta_{12} = \beta_{22} = \gamma, \\ 0 < \beta < 1, \\ \gamma > 1. \end{cases} \tag{3}$$

We seek an expression for the asymptotic probability distribution of response probabilities in terms of the numbers π, β, and γ.

The random variable η_n is defined recursively as follows:

$$\eta_1 = \begin{cases} \beta & \text{with prob } \pi, \\ \gamma & \text{with prob } (1 - \pi); \end{cases}$$

$$\eta_{n+1} = \begin{cases} \eta_n\beta & \text{with prob } \pi, \\ \eta_n\gamma & \text{with prob } (1 - \pi). \end{cases}$$

The random variable X_n is defined as follows:

$$X_n = \log \eta_n .$$

Then

$$X_{n+1} = \begin{cases} X_n + \log \beta & \text{with prob } \pi, \\ X_n + \log \gamma & \text{with prob } (1 - \pi). \end{cases} \tag{4}$$

It is clear from (4) and what has preceded that X_n is the sum of n independent identically distributed random variables Y_i defined by

$$Y_i = \begin{cases} \log \beta & \text{with prob } \pi, \\ \log \gamma & \text{with prob } (1 - \pi). \end{cases}$$

By the strong law of large numbers, with probability one as $n \to \infty$

$$\begin{aligned} X_n &\to \infty \quad \text{if} \quad \pi \log \beta + (1 - \pi) \log \gamma > 0, \\ X_n &\to -\infty \quad \text{if} \quad \pi \log \beta + (1 - \pi) \log \gamma < 0. \end{aligned} \tag{5}$$

Define now for any real number x

$$F_x(p_1) = \frac{p_1}{p_1 + e^x(1 - p_1)}. \tag{6}$$

Then $p_{n+1} = F_{X_n}(p_1)$ for the sequence of reinforcements η_n, where $X_n = \log \eta_n$. These results are utilized to prove the following theorem.

THEOREM 1. *Let* $c = \pi \log \beta + (1 - \pi) \log \gamma$. *Then with probability one*

$$p_\infty = \begin{cases} 0 & \text{if} \quad c > 0, \\ 1 & \text{if} \quad c < 0. \end{cases}$$

If $c = 0$, *then* p_n *oscillates between* 0 *and* 1, *so that with probability one*

$$\limsup p_n = 1$$

$$\liminf p_n = 0.$$

Despite this oscillation, there is a limiting distribution for p_n; *it is concentrated at* 0 *and* 1 *with equal probabilities* $\frac{1}{2}$.

PROOF. The results for $c > 0$ and $c < 0$ follow immediately from (5), (6), and the remark following. In case $c = 0$, note that $E(Y_i) = 0$. It is known [2] that the sums X_n are then recurrent; that is, they repeatedly take on values arbitrarily close to any possible value. In particular, X_n takes on repeatedly arbitrarily large and arbitrarily small values (with probability one), which upon recalling (6) proves the second statement. The third statement is a consequence of the central limit theorem, which implies that for any A, $\Pr(X_n > A)$ and $\Pr(X_n < -A)$ both converge to one-half as n increases. Again the assertion of the theorem follows from this fact and (6).

Two Theorems on Random Walks

The results of this section are special cases of those in [5]. However, the present approach has the advantages of simplicity and directness.

We have seen that the two-operator, noncontingent beta model gives rise to a Markov process on the real line such that from x the "moving particle" goes to $x + a$ or $x - b$ with (constant) probabilities φ and $1 - \varphi$. The contingent case leads to a similar process, except that the transition probabilities become functions of x. The four-operator model gives rise to a process with four possible transitions, from x to $x + a_i$, say, $i = 1, 2, 3, 4$. In this section some simple results on processes of these sorts will be obtained, in preparation for the study of the more general cases of the beta model. In the interest of clarity, only the two-operator case will be treated in full; the more general case can be handled in a similar way, but the details are cumbersome. Our approach was suggested by the work of Hodges and Rosenblatt [3].

Let $\{X_n\}$ be a real Markov process such that if $X_n = x$,

$$X_{n+1} = \begin{cases} x + a & \text{with prob } \varphi(x), \\ x - b & \text{with prob } [1 - \varphi(x)], \end{cases} \tag{9}$$

where $0 < a, b, \varphi(x), 1 - \varphi(x)$. Let $\{Y_n\}$ be another process of the same type (and with the same a and b) but with constants θ and $1 - \theta$ as the transition probabilities in place of $\varphi(x)$ and $1 - \varphi(x)$.

LEMMA. *If for all $x \geq M$, one has $\varphi(x) \geq \theta$, and if* $\Pr(Y_n \to +\infty) > 0$, *then* $\Pr(X_n \to +\infty) > 0$. *If, on the other hand, for $x \geq M$, $\varphi(x) \leq \theta$ and if* $\Pr(Y_n \to +\infty) = 0$, *then* $\Pr(X_n \to +\infty) = 0$.

PROOF. Let $\{\xi_n\}$ be a sequence of independent random variables, each uniformly distributed on $[0, 1]$. The $\{X_n\}$ process will be referred to $\{\xi_n\}$ by letting

$$X_{n+1} = \begin{cases} X_n + a & \text{if} \quad \xi_{n+1} \leq \varphi(X_n), \\ X_n - b & \text{otherwise.} \end{cases} \tag{10}$$

This does lead to the transition law (9) as may easily be seen. The $\{Y_n\}$ process can be linked to $\{X_n\}$ by referring it after the manner of (10) to the *same* sequence $\{\xi_n\}$, so that $Y_{n+1} = Y_n + a$ if and only if $\xi_{n+1} \leq \theta$.

Choose $Y_0 > M$. Whatever the value of X_0, since $\varphi(x) > 0$ there is positive probability that $X_m \geq Y_0$ for some m; therefore assume $X_0 \geq Y_0$. We now assert that for those sequences $\{Y_n\}$ with the property that $Y_n \geq M$ for all n, the inequality $X_n \geq Y_n$ is also valid for all n. This follows from our construction "linking" the processes, and the assumption that $\varphi(x) \geq \theta$ for $x \geq M$; the transition $X_{n+1} = X_n - b$ and $Y_{n+1} = Y_n + a$ is impossible, so $X_n - Y_n$ can only increase.

To complete the proof, note that since $\Pr(Y_n \to +\infty)$ is positive, so is $\Pr(Y_n \to +\infty, Y_n \geq M \text{ for all } n)$. But the event $Y_n \to +\infty$, $Y_n \geq M$ for all n may be considered as a set S in the sample space of the sequence $\{\xi_n\}$; S is a set of positive probability, and is contained in the set $X_n \to \infty$ since on S, $X_n \geq Y_n$ and $Y_n \to \infty$. Hence $\Pr(X_n \to +\infty) > 0$. The second part of the lemma is proved in a similar way, using the same construction linking $\{X_n\}$ and $\{Y_n\}$.

THEOREM 2. *Let $b/(a + b) = c$, and suppose that*

$$\lim_{x \to +\infty} \varphi(x) = \alpha \quad \textit{and} \quad \lim_{x \to -\infty} \varphi(x) = \beta \tag{11}$$

exist. Then if $\alpha < c$ and $\beta > c$,

$$\Pr(\limsup X_n = +\infty, \liminf X_n = -\infty) = 1 \quad (\{X_n\} \textit{ is recurrent}), \tag{12}$$

while if $\alpha < (>)\ c$ and $\beta < (>) c$, then

$$\Pr(X_n \to -\infty\ (+\infty)) = 1. \tag{13}$$

Finally, if $\alpha > c$ and $\beta < c$,

$$\Pr(X_n \to +\infty) = \delta, \qquad \Pr(X_n \to -\infty) = 1 - \delta \tag{14}$$

for some $0 < \delta < 1$.

PROOF. Suppose, for instance, that $\alpha < c$. Let $\{Y_n\}$ (as in the lemma) be a process with constant transition probabilities θ and $1 - \theta$ where $\alpha < \theta < c$. The $\{Y_n\}$ process may be regarded as sums of random variables

$$Y_n = Y_0 + \sum_{i=1}^{n} Z_i\,, \quad \text{where} \quad \Pr(Z_i = a) = \theta \quad \text{and} \tag{15}$$

$$\Pr(Z_i = -b) = 1 - \theta.$$

But $E(Z_i) = a\theta - b(1 - \theta) < 0$, since $\theta < c$; this implies that $\Pr(Y_n \to -\infty) = 1$ by the law of large numbers. From the lemma, $\Pr(X_n \to +\infty) = 0$.

Similarly, if $\alpha > c$ it follows that $\Pr(X_n \to +\infty) > 0$. Since the lemma also holds for convergence to $-\infty$ (with φ and θ replaced by $1 - \varphi$ and $1 - \theta$), we obtain in the same way that $\beta < c$ makes $\Pr(X_n \to -\infty) > 0$, while if $\beta > c$ this probability is zero.

Consider the case when $\alpha < c$ and $\beta < c$; there is then positive probability of absorption at $-\infty$, but not at $+\infty$. It is not hard to see that $X_n \to -\infty$ with probability one; the idea is roughly as follows. Since $X_n \nrightarrow +\infty$, we have $X_n \leq N$ infinitely often with probability arbitrarily close to 1 for some N. Now the probability that from or to the left of N the random walk goes and remains to the left of $N - M$ must be positive since $\Pr(X_n \to -\infty) > 0$. But in an infinite sequence of not necessarily independent trials, an event whose probability on each trial is bounded away from zero is certain to occur. Hence for any M, the random walk will eventually become and remain to the left of $N - M$, and therefore $X_n \to -\infty$ with probability arbitrarily close to 1 (and so equal to one). The other cases are similar; one can think of $\alpha > c$ or $\alpha < c$ as the conditions under which $+\infty$ is an absorbing or reflecting barrier, etc., and the process behaves accordingly.

The generalization to the four-operator case will now be described. Let $\{X_n\}$ be a real Markov process such that if $X_n = x$, then

$$X_{n+1} = x + a_i \quad \text{with prob } \varphi_i(x), \tag{17}$$

where $a_1, a_2 > 0 > a_3, a_4$ and $\varphi_i(x) > 0$. Suppose

$$\lim_{x \to +\infty} \varphi_i(x) = \alpha_i \quad \text{and} \quad \lim_{x \to -\infty} \varphi_i(x) = \beta_i \tag{18}$$

exist, and let

$$\mu_+ = \sum_{i=1}^{4} a_i \alpha_i \quad \text{and} \quad \mu_- = \sum_{i=1}^{4} a_i \beta_i\,.$$

By methods entirely similar to those used above, but rather more involved, it is possible to prove the following.

THEOREM 3. *For the process* $\{X_n\}$ *described above, if* $\mu_+ < 0$ *and* $\mu_- > 0$ *then* (12) *holds; if* $\mu_+ < (>) 0$ *and* $\mu_- < (>) 0$ *then* (13) *applies; while if* $\mu_+ > 0$ *and* $\mu_- < 0$, (14) *is valid.*

Contingent Reinforcement with Two Operators

If the probability of reinforcement depends only on the immediately preceding response (on the same trial), one has (*simple*) *contingent reinforcement.* Let $\Pr(E_1 \mid A_1) = \pi_1$ and $\Pr(E_1 \mid A_2) = \pi_2$, and let the two operators β and γ be specified as in (3). Using (6), define the random variable X_n recursively. (Note that $\log \gamma$ appears first, since $\log \gamma > 0$ and $\log \beta < 0$, in order most directly to apply Theorem 2.)

$$(19) \qquad X_{n+1} = \begin{cases} X_n + \log \gamma & \text{with prob } F_{X_n}(p_1)(1 - \pi_1) \\ & \qquad + (1 - F_{X_n}(p_1))(1 - \pi_2) = \varphi(X_n), \\ X_n + \log \beta & \text{with prob } [1 - \varphi(X_n)]. \end{cases}$$

Observe that

$$(20) \qquad \lim_{x \to +\infty} \varphi(x) = 1 - \pi_2 \quad \text{and} \quad \lim_{x \to -\infty} \varphi(x) = 1 - \pi_1 .$$

Combining (20) and Theorem 2, one then has immediately Theorem 4.

THEOREM 4. *For the contingent case of the two-operator model, let* $c = -\log \beta / \log (\gamma/\beta)$. *Then with probability one*

(i) *if* $1 - \pi_2 < c$ *and* $1 - \pi_1 > c$ *then*

$$\limsup_n p_n = 1 \quad \textit{and} \quad \liminf_n p_n = 0,$$

(ii) *if* $1 - \pi_2 < c$ *and* $1 - \pi_1 < c$ *then* $p_\infty = 1$,

(iii) *if* $1 - \pi_2 > c$ *and* $1 - \pi_1 > c$ *then* $p_\infty = 0$.

Moreover,

(iv) *if* $1 - \pi_2 > c$ *and* $1 - \pi_1 < c$ *then for some* δ *with* $0 < \delta < 1$

$$\Pr (p_n \to 1) = \delta, \qquad \Pr (p_n \to 0) = 1 - \delta.$$

The intuitive character of the distinction between the results expressed in (i) and (iv) of this theorem should be clear. If $1 - \pi_2 < c$ and $1 - \pi_1 > c$, then probability zero of an A_1 response and probability one of an A_1 response are both reflecting barriers, whereas if $1 - \pi_2 > c$ and $1 - \pi_1 < c$, they are both absorbing barriers.

It is also to be noticed that except when $1 - \pi_1 = c$ or $1 - \pi_2 = c$, Theorem 4 covers all values of β, γ, π_1, and π_2 for the contingent case. It can be shown [5] by deeper methods that if $1 - \pi_1 = c$ (or $1 - \pi_2 = c$) then probability one (respectively zero) of an A_1 response is again a reflecting barrier. These results agree with those given by Luce ([7], p. 124) and in addition settle most of the open questions in his Table 6. Detailed comparison is tedious because his classification of cases differs considerably from ours as given in the above theorem.

Contingent Reinforcement with Four Operators

We want finally to apply Theorem 3 to the contingent case of the general four-operator model formulated in (1). Analogous to (19),

$$X_{n+1} = \begin{cases} X_n + \log \beta_{22} & \text{with prob } (1 - \pi_2)(1 - F_{X_n}(p_1)) = \varphi_{22}(X_n), \\ X_n + \log \beta_{12} & \text{with prob } (1 - \pi_1)F_{X_n}(p_1) = \varphi_{12}(X_n), \\ X_n + \log \beta_{21} & \text{with prob } \pi_2(1 - F_{X_n}(p_1)) = \varphi_{21}(X_n), \\ X_n + \log \beta_{11} & \text{with prob } \pi_1 F_{X_n}(p_1) = \varphi_{11}(X_n). \end{cases} \tag{21}$$

Also,

$$\begin{cases} \lim_{x \to +\infty} \varphi_{22}(x) = 1 - \pi_2 , & \lim_{x \to -\infty} \varphi_{22} = 0, \\ \lim_{x \to +\infty} \varphi_{12}(x) = 0, & \lim_{x \to -\infty} \varphi_{12}(x) = 1 - \pi_1 , \\ \lim_{x \to +\infty} \varphi_{21}(x) = \pi_2 , & \lim_{x \to -\infty} \varphi_{21}(x) = 0, \\ \lim_{x \to +\infty} \varphi_{11}(x) = 0, & \lim_{x \to -\infty} \varphi_{11}(x) = \pi_1 . \end{cases} \tag{22}$$

Then

$$\mu_+ = \sum_{j,k} \log \beta_{jk} \lim_{x \to +\infty} \varphi_{jk}(x) = \pi_2 \log \beta_{21} + (1 - \pi_2) \log \beta_{22} , \tag{23}$$

and

$$\mu_- = \sum_{j,k} \log \beta_{jk} \lim_{x \to -\infty} \varphi_{jk}(x) = \pi_1 \log \beta_{11} + (1 - \pi_1) \log \beta_{12} . \tag{24}$$

To apply Theorem 3 one also assumes that β_{22} , $\beta_{12} > 1 > \beta_{21}$, $\beta_{11} > 0$. On this assumption, and utilizing (23) and (24), we infer Theorem 5.

THEOREM 5. *For the contingent case of the four-operator model, with probability one*

(i) *if* $\mu_+ < 0$ *and* $\mu_- > 0$ *then* $\limsup_n p_n = 1$ *and* $\liminf_n p_n = 0$,

(ii) *if* $\mu_+ < 0$ *and* $\mu_- < 0$ *then* $p_\infty = 1$,

(iii) *if* $\mu_+ > 0$ *and* $\mu_- > 0$ *then* $p_\infty = 0$;

and if $\mu_+ > 0$ *and* $\mu_- < 0$, *then for some* δ *with* $0 < \delta < 1$

(iv) $\Pr(p_n \to 1) = \delta$, $\Pr(p_n \to 0) = 1 - \delta$.

Specialization of this theorem to cover the noncontingent case is immediate.

REFERENCES

[1] Bush, R. R., Galanter, E., and Luce, R. D. Tests of the "beta model." In R. R. Bush and W. K. Estes (Eds.), *Studies in mathematical learning theory*. Stanford: Stanford Univ. Press, 1959. Ch. 18.

[2] Chung, K. L. and Fuchs, W. H. J. On the distribution of values of sums of random variables. *Mem. Amer. Math. Soc.*, 1951, **6**, 1–12.

[3] Hodges, J. L. and Rosenblatt, M. Recurrence time moments in random walks. *Pac. J. Math.*, 1953, **3**, 127–136.

[4] Karlin, S. Some random walks arising in learning models I. *Pac. J. Math.*, 1953, **3**, 725–756.

[5] Lamperti, J. Criteria for the recurrence or transience of stochastic processes I. *J. math. Anal. Applications*, (in press).

[6] Lamperti, J. and Suppes, P. Chains of infinite order and their application to learning theory. *Pac. J. Math.*, 1959, **9**, 739–754.

[7] Luce, R. D. *Individual choice behavior*. New York: Wiley, 1959.

Manuscript received 4/27/59

Revised manuscript received 11/10/59

PSYCHOMETRIKA—VOL. 25, NO. 3
SEPTEMBER, 1960

PERCENTAGE POINTS OF WILKS' L_{mvc} AND L_{vc} CRITERIA*

J. ROY AND V. K. MURTHY

UNIVERSITY OF NORTH CAROLINA

Likelihood ratio tests have been proposed by Wilks for testing the hypothesis of equal means, variances, and covariances (H_{mvc}) and the hypothesis of equal variances and covariances (H_{vc}) in a p-variate normal distribution. Using exact distributions of the appropriate likelihood ratio statistics, tables of the .05 and .01 points of these distributions are constructed for $p = 4, 5, 6, 7$ and sample size $n = 25\,(5)\,60\,(10)\,100$. A correction factor is recommended for larger n. Two numerical examples illustrate use of the tables. A nonparametric test is proposed for H_{mvc} when the multivariate parent population is known to be non-normal.

In connection with a p-variate normal population Wilks [5] proposed the following hypotheses: (i) H_{mvc} : that the means are equal, the variances are equal, and the covariances are equal; and (ii) H_{vc} : that the variances are equal and the covariances are equal. These hypotheses are of great importance in psychometrics, especially in the theory of mental tests [1]. The concept of parallel tests, for instance, leads to an examination of the hypothesis H_{mvc} .

In a random sample of size n from a p-variate normal population, let $x_{i\lambda}$ denote the value of the ith variate for the λth individual $i = 1, 2, \cdots, p$, $\lambda = 1, 2, \cdots, n$. Let

$$\bar{x}_i = \frac{1}{n} \sum_{\lambda=1}^{n} x_{i\lambda}, \qquad S_{ij} = \frac{1}{n} \sum_{\lambda=1}^{n} (x_{i\lambda} - \bar{x}_i)(x_{j\lambda} - \bar{x}_j),$$

$$T = \frac{1}{p} \sum_{i=1}^{p} S_{ii}, \qquad U = \frac{1}{p(p-1)} \sum_{i \neq} \sum_{j=1}^{p} S_{ij}, \qquad \bar{x} = \frac{1}{p} \sum_{i=1}^{p} \bar{x}_i .$$

Wilks [5] showed that the likelihood ratio principle gives the following test procedure. At the level of significance α, $0 < \alpha < 1$, reject the hypothesis H_{mvc} (H_{vc}) if the test criterion $L_{mvc}(L_{vc})$ falls short of the constant L_α , the lower 100α percentage point of the distribution of $L_{mvc}(L_{vc})$; otherwise accept the hypothesis $H_{mvc}(H_{vc})$. The test criteria are defined by

$$(1) \qquad L_{mvc} = \frac{| S_{ij} |}{[T + (p-1)U]\left[T - U + \dfrac{1}{p-1} \displaystyle\sum_{i=1}^{p} (\bar{x}_i - \bar{x})^2\right]^{p-1}},$$

*This research was supported partly by the Office of Naval Research under Contract No. Nonr-855(06) and partly by the United States Air Force through the Air Force Office of Scientific Research of the Air Research and Development Command, under Contract No. 18(600)-83. Reproduction in whole or in part for any purpose of the United States Government is permitted.

and

$$L_{vc} = \frac{|S_{ij}|}{[T + (p-1)U](T-U)^{p-1}}, \tag{2}$$

respectively. Wilks [5] computed the moments of these criteria and showed that asymptotically, for large n, the transforms $-n \log_e L_{mvc}$ and $-n \log_e L_{vc}$ are distributed as chi square with $\frac{1}{2}p(p+3) - 3$ and $\frac{1}{2}p(p+1) - 2$ degrees of freedom, respectively. He also worked out the exact distribution of these criteria in the cases $p = 2$ and $p = 3$.

The chi-square approximation, however, is not good enough for moderately large values of n as it considerably overestimates significance. Approximation by a beta variable has been suggested by Wilks and Tukey [6], but since this requires double interpolation in available tables, it is not very convenient. Varma [4] derived the exact distribution in a series form, which, however is not easy to tackle.

In view of the special importance of these criteria in psychometric work, it seems useful to derive the exact distribution in a convenient form and compute the percentage points for ready use. In this paper an asymptotic series expansion derived by one of the authors [2, 3] is used to evaluate .05 and .01 points of the distribution of L_{mvc} and L_{vc} criteria for $p = 4, 5, 6, 7$ and $n = 25\ (5)\ 60\ (10)\ 100$. For higher values of n a correction factor a is provided such that to a high degree of accuracy $-(n - a)\log_e L_{mvc}$ or $-(n - a)\log_e L_{vc}$ is distributed as chi square. The use of the tables is illustrated with two numerical examples. A simple nonparametric alternative procedure is suggested for testing a generalization of the H_{mvc} hypothesis.

The Method for Computing the Percentage Points

It has been shown [2, 3] that for a properly chosen constant a, writing

$$X = -N \log_e L, \tag{3}$$

where

$$N = n - a \tag{4}$$

and L is Wilks' L_{mvc} or L_{vc} criterion based on a sample of size n from a p-variate normal population, the probability density function $f_N(x)$ of X can be expressed as

$$f_N(x) = C_N\left[p_r(x) + \frac{a_2}{N^2}p_{r+4}(x) + \frac{a_3}{N^3}p_{r+6}(x) + \frac{a_4}{N^4}p_{r+8}(x) + \cdots\right], \tag{5}$$

where $p_r(x)$ is the chi-square density function with r degrees of freedom,

$$p_r(x) = \frac{1}{2^{r/2}\Gamma(r/2)}e^{-x/2}x^{(r/2)-1}. \tag{6}$$

C_N is a function of N that can be expanded in an asymptotic series

$$1/C_N = 1 + \frac{a_2}{N^2} + \frac{a_3}{N^3} + \frac{a_4}{N^4} + \cdots, \tag{7}$$

and r, a, a_2, a_3, a_4 are constants independent of N. The values of these constants are tabulated in Table 1 and Table 2 for $p = 4, 5, 6, 7$, and 8.

TABLE 1

Values of the Constants in the Asymptotic Expansion of the Distribution of Wilks' L_{mvc} Criterion

p	r	a	a_2	a_3	a_4
4	11	2.21718	3.48140	0.55608	9.33616
5	17	2.48530	8.24908	2.49784	50.72240
6	24	2.77500	16.29624	7.05128	180.95904
7	32	3.07638	28.78664	15.92480	530.00624
8	41	3.38502	47.05200	31.38352	1352.19984

TABLE 2

Values of the Constants in the Asymptotic Expansion of the Distribution of Wilks' L_{vc} Criterion

p	r	a	a_2	a_3	a_4
4	8	2.73611	1.47184	0.26324	60.83198
5	13	3.01923	4.24880	1.33136	187.00837
6	19	3.32105	9.41040	4.06681	456.04610
7	26	3.63248	17.95537	9.73818	978.32878
8	34	3.94958	31.04982	20.09536	1925.08274

For any given number α, $0 < \alpha < 1$, to compute L_α the lower 100 α percentage point of the distribution L proceed as follows.

Obviously,

$$L_\alpha = \exp(-X_\alpha/N), \tag{8}$$

where X_α is defined by

$$\text{Prob}\,(X > X_\alpha) = \alpha.$$

Let x_0 be the upper 100 α percentage point of the chi-square distribution with r degrees of freedom, that is

$$Q_r(x_0) = \alpha, \tag{9}$$

where

$$Q_r(x) = \int_x^\infty p_r(\xi)\, d\xi. \tag{10}$$

Since for large N, $C_N \sim 1$ and the second and succeeding terms in (5) are negligible, as a first approximation

$$X_\alpha \doteq x_0 \,. \tag{11}$$

Let δ be the additive correction to be made to x_0 so that

$$X_\alpha = x_0 + \delta. \tag{12}$$

Retaining only the first two terms,

$$\alpha\left(1 + \frac{a^2}{N^2}\right) = Q_r(x_0 + \delta) + \frac{a^2}{N^2} Q_{r+4}(x_0 + \delta). \tag{13}$$

Expanding the right-hand side in a Taylor series about x_0 and neglecting δ^2 and higher powers of δ,

$$\delta_0 = \frac{2\left(\frac{x_0^2}{r(r+2)} + \frac{x_0}{r}\right)\frac{a_2}{N^2}}{1 + \frac{x_0^2}{r(r+2)}\cdot\frac{a_2}{N^2}} \tag{14}$$

as an approximation for δ. As a second approximation to X_α ,

$$x_1 = x_0 + \delta_0 \,. \tag{15}$$

Now, from (5) one has the asymptotic expansion

$$\begin{aligned} \text{prob}\,(X \geq x) &= Q_r(x) + \frac{a_2}{N^2}\,[Q_{r+4}(x) - Q_r(x)] \\ &\quad + \frac{a_3}{N^3}\,[Q_{r+6}(x) - Q_r(x)] + \left[\frac{a_4}{N^4}\,[Q_{r+8}(x) - Q_r(x)] \right. \\ &\quad \left. - \frac{a_2}{N^2}\,[Q_{r+4}(x) - Q_r(x)]\right] + O\!\left(\frac{1}{N^5}\right). \end{aligned} \tag{16}$$

The required percentage point is in the neighborhood of x_1 and can be evaluated by inverse interpolation by first tabulating Prob $(X \geq x)$ for several values of x around x_1 . As it happens, however, further corrections to x_1 become necessary only for very small values of N.

Tables of Percentage Points

The .05 and .01 points of the distribution of Wilks' L_{mvc} and L_{vc} criteria are given to four decimal places for $p = 4, 5, 6$, and 7 in Table 3 and Table 4

respectively for $n = 25$ (5) 60 (10) 100. For n greater than 100, $-(n-a)$ $\log_e L_{mvc}$ can safely be used as chi square with r degrees of freedom. We have not extended the tables below $n = 25$ first because the asymptotic expansion (16) with only four terms is not good enough for n smaller than 25, and second because a sample of size less than 25 is not to be recommended in multivariate work with four or more variates.

TABLE 3

.05 and .01 Points of L_{mvc} Criterion

	.05 point				.01 point			
n\p	4	5	6	7	4	5	6	7
25	.4206	.2920	.1923	.1196	.3366	.2251	.1427	.0854
30	.4918	.3658	.2623	.1781	.4098	.2957	.2048	.1356
35	.5482	.4273	.3229	.2339	.4698	.3570	.2623	.1859
40	.5937	.4787	.3759	.2852	.5193	.4098	.3143	.2338
45	.6311	.5222	.4220	.3314	.5607	.4552	.3606	.2783
50	.6623	.5592	.4625	.3730	.5958	.4946	.4019	.3191
55	.6887	.5911	.4979	.4102	.6258	.5290	.4387	.3654
60	.7113	.6187	.5292	.4437	.6518	.5591	.4715	.3903
70	.7480	.6645	.5815	.5011	.6943	.6096	.5274	.4494
80	.7763	.7005	.6239	.5483	.7276	.6498	.5730	.4985
90	.7992	.7296	.6586	.5876	.7545	.6826	.6108	.5403
100	.8177	.7536	.6875	.6208	.7765	.7099	.6426	.5757

TABLE 4

.05 and .01 Points of L_{vc} Criterion

	.05 point				.01 point			
n\p	4	5	6	7	4	5	6	7
25	.5129	.3768	.2473	.1601	.4209	.2985	.1866	.1163
30	.5773	.4490	.3219	.2273	.4908	.3709	.2563	.1756
35	.6271	.5071	.3853	.2883	.5464	.4313	.3181	.2322
40	.6666	.5546	.4390	.3424	.5913	.4819	.3721	.2841
45	.7002	.5941	.4847	.3899	.6284	.5248	.4191	.3310
50	.7251	.6273	.5239	.4318	.6594	.5613	.4601	.3731
55	.7473	.6555	.5578	.4686	.6857	.5928	.4961	.4108
60	.7663	.6799	.5873	.5013	.7083	.6201	.5278	.4446
70	.7967	.7196	.6362	.5564	.7450	.6653	.5810	.5024
80	.8202	.7506	.6748	.6008	.7736	.7011	.6236	.5499
90	.8394	.7755	.7062	.6374	.7964	.7299	.6586	.5894
100	.8560	.7959	.7321	.6679	.8151	.7538	.6877	.6226

Illustrative Example

Table 5 gives means, variances, and covariances for scores on four tests for a sample of 50 examinees.

TABLE 5

Means, Variances, and Covariances for Four Tests, n = 50

Tests	Means	Dispersion matrix			
		A	B	C	D
A	14.9048	25.0704	12.4363	11.7257	20.7510
B	15.4841		28.2021	9.2281	11.9732
C	14.4444			22.7390	12.0692
D	14.3810				21.8707

(a) Can the tests be regarded as parallel? (b) If not, would additive corrections applied to the means make the tests parallel?

Tests are said to be parallel [1] if test scores obtained in the population of examinees have equal means, equal variances, and equal covariances. To answer question (a) the appropriate hypothesis to be tested is H_{mvc}, and to answer (b) the appropriate hypothesis to be tested is H_{vc}. The numerical procedure for carrying out these tests is given below.

In this case, $p = 4$ and

$$|S_{ij}| = 39750.5, \qquad T = 24.47055, \qquad U = 13.03058,$$

$$\sum (\bar{x}_i - \bar{x})^2 = 0.78094, \; T + (p-1)U = 63.56229, \; T - U = 11.43997.$$

Thus

$$L_{mvc} = \frac{39750.5}{(63.56229)(11.43997 + 0.26031)^3}$$

$$= \frac{39750.5}{104040.6} = 0.3821.$$

The .01 point of L_{mvc} for $p = 4$ and $n = 50$ is 0.5958 and therefore the hypothesis H_{mvc} must be rejected at the .01 level of significance.

The L_{vc} statistic turns out to be

$$L_{vc} = \frac{39750.5}{(63.56229)(11.43997)^3}$$

$$= \frac{39750.5}{95164.3} = 0.4177.$$

The .01 point of L_{vc} for $p = 4$ and $n = 50$ is 0.6594 and therefore the hypothesis H_{vc} has to be rejected at the .01 level. Therefore, the tests are not parallel and even additive corrections in the means would not make them so.

A Nonparametric Test for Symmetry

The hypothesis H_{mvc} for a multivariate normal population is equivalent to the hypothesis that the joint distribution function is symmetric, that is, invariant under a permutation of the variates. However, the statistic L_{mvc} is not appropriate for testing symmetry of a multivariate distribution which is not definitely known to be normal. A simple nonparametric test for symmetry appropriate for any continuous multivariate distribution is proposed here.

Let $(X_1, X_2, \cdots, X_p)$ be a randomly selected observation from a continuous p-variate distribution and let $i \equiv (i_1, i_2, \cdots, i_p)$ be a permutaion of the integers $(1, 2, \cdots, p)$. We shall say the observation is of "type i" if

$$X_{i_1} < X_{i_2} < \cdots < X_{i_p}$$

holds. Thus with probability 1 every observation belongs to one and only one of the $p!$ types. If the hypothesis of symmetry is true, all these types are equally likely, the probability that a random observation is of any particular type being $1/p!$. If in n observations, n_i are of type i, compute

$$\chi^2 = \frac{p!}{n} \sum \left(n_i - \frac{n}{p!}\right)^2 = \frac{p!}{n} \sum n_i^2 - n,$$

the summation being over all types. If this exceeds the upper α point of the chi-square statistic with $(p! - 1)$ degrees of freedom, the hypothesis of symmetry is rejected at level of significance α.

TABLE 6

Scores on the Tests A, B, and C

A	B	C	A	B	C	A	B	C
56	40	46	50	46	23	42	36	0
34	57	48	31	45	20	39	55	10
32	47	38	43	37	32	37	58	40
55	24	32	59	43	58	28	24	29
37	63	59	38	48	14	37	52	40
32	40	7	29	36	38	62	45	50
33	58	34	27	53	18	24	29	13
62	74	58	38	35	22	32	35	39
28	42	36	40	61	12	47	36	15
41	60	16	41	42	26	45	46	24
20	35	7	46	62	32	52	43	44
47	39	24	55	54	24	65	72	84
33	53	54	52	43	15	31	49	36
44	40	31	46	38	17	54	62	64
41	42	28	48	52	61	51	48	53
28	40	42	59	52	63	40	36	42
47	50	64	55	47	56			

Illustrative Example

In Table 6 appear scores on tests A, B, and C, for a random sample of 50 examinees. Test the hypothesis of symmetry.

In this example the $3! = 6$ types of permutations are $X_1 < X_2 < X_3$; $X_1 < X_3 < X_2$; $X_2 < X_1 < X_3$; $X_2 < X_3 < X_1$; $X_3 < X_1 < X_2$; and $X_3 < X_2 < X_1$; let us call them permutations of type 1, 2, 3, 4, 5, and 6 respectively. Table 7 gives the observed frequencies of the six types; the expected frequency of each type when the hypothesis of symmetry is true is 8.3333.

TABLE 7

Observed Frequency Distribution

Type i	Frequency n_i
1	8
2	8
3	5
4	5
5	14
6	10
Total	50

Here $p = 3$, $n = 50$ and

$$\chi^2 = \frac{p!}{n} \sum n_i^2 - n = \frac{6}{50}(474) - 50$$
$$= 6.88.$$

The .05 point of χ^2 with $p! - 1 = 5$ degrees of freedom is 11.07; thus there is no evidence for rejecting the hypothesis of symmetry.

REFERENCES

[1] Gulliksen, H. *Theory of mental tests.* New York: Wiley, 1951.

[2] Roy, J. The distribution of certain likelihood criteria useful in multivariate analysis. *Bull. int. statist. Inst.*, 1951, **33**, 219–230.

[3] Roy, J. Tests of independence and symmetry in multivariate normal populations. Unpublished thesis, Calcutta Univ.

[4] Varma, K. B. On the exact distribution of Wilks' L_{mvc} and L_{vc} criteria. *Bull. int. statist. Inst.*, 1951, **33**, 181–214.

[5] Wilks, S. S. Sample criteria for testing equality of means, equality of variances, and equality of covariances in a normal multivariate distribution. *Ann. math. Statist.*, 1946, **17**, 257–281.

[6] Wilks, S. S. and Tukey, J. W. Approximation of the distribution of the product of Beta variables by a single Beta variable. *Ann. math. Statist.*, 1946, **17**, 318–324.

Manuscript received 6/25/58
Revised manuscript received 2/9/60

PSYCHOMETRIKA—VOL. 25, NO. 3
SEPTEMBER, 1960

MODELS FOR CHOICE-REACTION TIME

MERVYN STONE

MEDICAL RESEARCH COUNCIL*

In the two-choice situation, the Wald sequential probability ratio decision procedure is applied to relate the mean and variance of the decision times, for each alternative separately, to the error rates and the ratio of the frequencies of presentation of the alternatives. For situations involving more than two choices, a fixed sample decision procedure (selection of the alternative with highest likelihood) is examined, and the relation is found between the decision time (or size of sample), the error rate, and the number of alternatives.

This paper develops to the point of usefulness several mathematical models for choice-reaction time. The working details are confined to appendices and only definitions and results appear in the text. It is hoped that this method of presentation will assist the reader in making a quick "calculated-observed" analysis of the data he may have. The choice of models is made mainly by analogy with statistical decision procedures, but no model is presented which is psychologically unreasonable. Also no comparisons are made with experimental data for several reasons: (i) the paucity of available data means that the field should be kept open to avoid premature rejections; (ii) published data are often summarized in directions orthogonal to our interests; (iii) for the most powerful discrimination, experiments will need to be designed with specific models in mind.

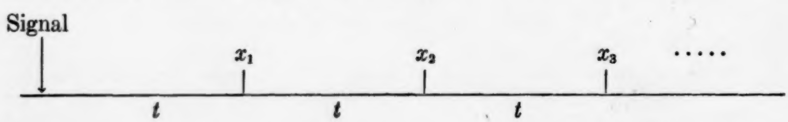
The models are envisaged as applying to the situation in which the subject (S) is given a time-stationary stimulus or signal and is required to identify some attribute of the signal and make an appropriate reaction. The signal remains present until the reaction is made. S is presented with signal after signal and the successive attributes form a random sequence; that is, for a given run of signals, the attributes of different signals are mutually independent and their probabilities of presentation do not change with time. The models assume that S has a settled mode of response. They will be hydrodynamic in the following sense. At the onset of each signal, a stream of information about the signal flows at a uniform rate into S. After a certain time, the input time, the front of this stream reaches S's decision taking mechanism or "computer." After a further time, the decision time, S makes a response. The time taken for the response to be recorded will be called the motor time. Thus the choice-reaction time is made up of three components:

*Applied Psychology Research Unit, 15 Chaucer Road, Cambridge, England.

the input time, T_i ; the decision time, T_d ; the motor time, T_m . The models apply to T_d , which will be related to the environmental variables (the number of signals and their frequencies of presentation) and the rate at which S makes incorrect responses. By concentrating on T_d in this way, it is not implied that T_i and T_m are necessarily independent of these factors.

Likelihood Ratio Models for the Two-Choice Situation

It is assumed that the subject knows when the signal (either s_0 or s_1 , say) commences; that is, he knows when to start examining the stream of information arriving at the computer. (This stream is "noisy" until the stream from the signal is added to it.) This assumption holds in the self-paced condition and also when some preparatory warning signal is given. It is supposed that there is some overlap in the information; that is, some patterns of information may arise from either s_0 or s_1 . If there is no uncertainty in this sense, there is no need for a statistical computer. The uncertainty may arise from the external situation, from noise added at the input stage, or from both sources. We will suppose that the information on which S's computer operates is equivalent to a series of independent random variables at short time intervals t and that each random variable has the (stationary) distribution of a random variable x (dependent on which signal has occurred) until the response is made.



Let $p_0(x)$ and $p_1(x)$ be the probabilities of x when the signal is s_0 and s_1 , respectively. If the x's are instantaneous samples of an almost continuous stream of information then the assumption of independence implies zero auto-correlation between parts of the stream not less than time t apart. If the x's are integrals of the stream over the successive intervals, then the assumption requires zero auto-correlation for all time lags (or at least for those not small compared with t). Suppose the computer transforms each x to a quantity $c(x)$ which is then stored in an adder.

Sequential Case

The computer makes a running total of $c(x_1)$, $c(x_2)$, $\cdots$. Constant log A and log B with $A > B$ are preselected so that S decides for s_0 (and makes the appropriate motor action) as soon as the total falls below log B, provided the total has not previously exceeded log A when the decision would have been made for s_1 . (The odd way of expressing the constants facilitates later

references.) If the decision is made at the nth sample $T_d = nt$. The theory of the sequential probability ratio test [1] shows that the optimum choice of the function $c(x)$ is

$$(1) \qquad c(x) = \log p_1(x) - \log p_0(x).$$

Such a function implies that S is familiar with the probability distributions $p_0(x)$ and $p_1(x)$. Such familiarity may be the result of a process of learning, provided S has performed many trials of the discrimination task and is given knowledge of results. S's computer may be thought of as exploratory, trying out different $c(x)$'s until the optimal one is found. However it is conceivable that the distributions can be deduced by S from the structure of the situation and then imposed on his computer. The optimality of (1) is stated by Wald [1] in the following terms: let $\bar{n}_0$, $\bar{n}_1$ be the averages of the number of samples necessary for decision when the signals presented are s_0, s_1, respectively. If $\bar{n}_0^*$, $\bar{n}_1^*$ are the averages for any other decision procedure based on x_1, x_2, etc., with smaller probabilities of incorrect response to s_0 and s_1, then $\bar{n}_0^* \geq \bar{n}_0$ and $n_1^* \geq \bar{n}_1$. It is possible that this form of optimality does not appeal to S, who may have to be trained to use it by suitable reward.

Before testing the model, it must be remembered that it is T which is measured and not T_d. Even so, a test is available which requires only the following assumption. Consider trials leading to a decision for s_0. The assumption is, given the value of T_d, that the distribution of $T_i + T_m$ is the same whether the decision is right or wrong. (The same assumption is made for decisions for s_1.) This does not exclude the possibility that $T_i + T_m$ and T_d be correlated. The length of time, T_i, may affect the uncertainty in the information presented to the computer and therefore may affect T_d; alternatively, if T_d is long, T_m may be deliberately shortened. However, it does assume that T_m cannot be influenced by information processed since the initiation of the motor action. In Appendix 1 it is shown that, with mild restrictions on $p_0(x)$ and $p_1(x)$, the distribution of the n's, and therefore of the T_d's, leading to a decision for s_0 (or of those leading to s_1) is the same whether the decisions are correct or incorrect. With the above assumption, this implies that the same result should hold for a comparison of the correct and incorrect T's leading to s_0 (and for a comparison of those leading to s_1). This provides the basis of a reasonable test of the model. However, a fair proportion of errors would be needed to give a powerful test.

Without making assumptions about $p_0(x)$ and $p_1(x)$, it is difficult to think of more ways of examining the validity of the model. Since x is an intervening variable without operational definition, it would clearly be unwise to assume much about $p_0(x)$ and $p_1(x)$. However, there is one assumption, called the "condition of symmetry," which in some discrimination

tasks may be reasonable. This is that the distribution of $p_1(x)/p_0(x)$, when x is distributed according to $p_0(x)$, is identical with that of $p_0(x)/p_1(x)$, when x is distributed according to $p_1(x)$. It is shown in Appendix 2 that, if this condition holds,

$$\bar{n}_1/\bar{n}_0 = J(\beta, \alpha)/J(\alpha, \beta); \tag{2}$$

$$\begin{aligned} J(\alpha, \beta)v_1 - J(\beta, \alpha)v_0 \\ = 4[J(\beta, \alpha)\alpha(1-\alpha)\bar{n}_1^2 - J(\alpha, \beta)\beta(1-\beta)\bar{n}_0^2]/(1-\alpha-\beta)^2, \end{aligned} \tag{3}$$

where α and β are the probabilities of incorrect response to a single s_0 and s_1, respectively, v_i is the variance of the sample sizes when s_i is presented, and

$$J(\alpha, \beta) = \alpha \log [\alpha/(1-\beta)] + (1-\alpha) \log [(1-\alpha)/\beta].$$

If it is feasible to estimate T_d directly for each trial by eliminating $T_i + T_m$ from T, then (2) and (3) imply

$$\bar{T}_{d1}/\bar{T}_{d0} = J(\beta, \alpha)/J(\alpha, \beta), \tag{4}$$

$$\begin{aligned} J(\alpha, \beta) \text{ var } T_{d1} - J(\beta, \alpha) \text{ var } T_{d0} \\ = 4[J(\beta, \alpha)\alpha(1-\alpha)\bar{T}_{d1}^2 - J(\alpha, \beta)\beta(1-\beta)\bar{T}_{d0}^2]/(1-\alpha-\beta)^2. \end{aligned} \tag{5}$$

Equations (4) and (5) are most relevant if S can be persuaded to achieve different (α, β) combinations without changing the distributions $p_0(x)$ and $p_1(x)$. When $\alpha = \beta$, then $\bar{n}_0 = \bar{n}_1$ and $v_0 = v_1$; with the assumptions that $T_i + T_m$ is (i) uncorrelated with T_d and (ii) independent of the signal presented, this implies equality of means and variances of reaction times to the signals. So, for the latter special case, it is not necessary to measure T_d.

For the "condition of symmetry" it is sufficient that, with x represented as a number, $p_0(x) = p_1(x - d)$ for some number d with $p_0(x)$ symmetrical about its mean. This might occur when s_0, s_1 are signals which are close together on some scale and the error added to the signals to make x has the same distribution for each signal. Symmetry would not be expected in absolute threshold discriminations or in the discrimination of widely different colors in a color-noisy background. Another sufficient condition is that x be bivariate, $[x(1), x(2)]$, the probabilities under s_0 obtained from those under s_1 by interchanging $x(1)$ and $x(2)$. For instance, $x(1)$ and $x(2)$ may be the inputs on two noisy channels and s_0 consists of stimulation of the first while s_1 consists of stimulation of the second.

A further prediction of the model for the symmetrical case can be made when S is persuaded by a suitable reward to give equal weight to errors to s_0 and s_1, that is to minimize his unconditional error probability, by adjustment of the constants A and B in his computer. If p_0 is the frequency of presentation of s_0 then the error probability is $p_0\,\alpha + (1 - p_0)\beta$ or e, say, and the average decision time is $p_0\bar{T}_{d0} + (1 - p_0)\bar{T}_{d1}$ or $\bar{T}_d$, say. It is shown

in Appendix 3 that, provided $10e < p_0 < 1 - 10e$, the minimization results in the following relation between $\bar{T}_d$, e and p_0:

$$\bar{T}_d \propto [J(e, 1 - e) - J(p_0, 1 - p_0)].$$

The Non-Sequential Fixed-Sample Case

If S has an incentive to react quickly and correctly, then the advantage of the sequential decision procedure is that those discriminations which by chance happen to be easy are made quickly and time is saved. However it is possible that S may adopt a different, less efficient strategy—which is to fix T_d for all trials at a value which will give a certain accepted error rate. Let the sample size corresponding to this decision time be n. The likelihood ratio procedures are as follows: decide for s_0 if $c(x_1) + \cdots + c(x_n) < \log C$; decide for s_1 if $c(x_1) + \cdots + c(x_n) \geq \log C$; $c(x) = \log p_1(x) - \log p_0(x)$ and $C > 0$. These procedures are optimal in the sense that, if any other procedure based on $x_1, \cdots, x_n$ is used, there exists one of the likelihood ratio procedures with smaller error probabilities. It was remarkable that in the sequential case useful predictions were obtainable under mild restrictions on $p_0(x)$ and $p_1(x)$. Unfortunately this does not hold for the fixed-sample case, making more difficult the problem of testing whether such a model holds.

If there is no input storage, it is possible that the results of the self-imposed strategy just outlined are equivalent to those obtainable when the experimenter himself cuts off the signals after an exposure time T_d. But this is the type of situation considered by Peterson and Birdsall [2]. The emphasis of these authors is mainly on the external parameters (such as energy) rather than on any supposed intervening variable. They define a set of physical situations for auditory discrimination in terms of a parameter d, which is equivalent to the difference between the means of two normal populations with unit variance. (For, in the cases considered, it happens that the logarithm of the likelihood ratio of the actual physical random variables for the two alternatives is normally distributed with equality of variance under the two alternatives.) This parameter sets a limit to the various performances (error probabilities to s_0 and s_1) of any discriminator using the whole of the physical information. It therefore sets an upper bound on the performance of S who can only use less than the whole. In [2] the authors make the assumption that the information on the basis of which S makes his discrimination nevertheless gives normality of logarithm of the likelihood ratio. They examine data to see whether S is producing error frequencies that lie on a curve defined by a d greater than that in the external situation.

More than Two Alternatives

For m alternatives there are m probability distributions for the intervening variable x (which may be multivariate); that is, signal s_i induces an

x with the probability distribution $p_i(x)$ for $i = 1, \cdots, m$. We will consider the consequences of a fixed-sample decision procedure based on $x_1, \cdots, x_n$, where n is fixed.

If the signals are presented independently with probabilities $p_1, \cdots, p_m$ (adding to unity) and if $\alpha_i(\mathfrak{D})$ is the probability of error to signal s_i when the decision procedure $\mathfrak{D}$ (based on $x_1, \cdots, x_n$) is used, then the probability of error to a single presentation is

$$e = \sum_1^m p_i \alpha_i(\mathfrak{D}).$$

It is shown in Appendix 4 that the $\mathfrak{D}$ minimizing e is that which effectively selects the signal with maximum posterior probability. In this section, this minimum e will be related to n (or T_d/t) and m when distributions are normal. However in the validation of the model it might be necessary to supplement T_d with a time $T_{\bar{d}}$, representing the time the computer requires to examine the m posterior probabilities to decide which is the largest. For, although it might be reasonable to suppose that $T_i + T_m$ is independent of m, one would expect $T_{\bar{d}}$ to vary with m. The simplest model for $T_{\bar{d}}$ would be to suppose that $T_{\bar{d}} = (m - 1)t'$, where t' is the time necessary to compare any two of the probabilities and decide which is the larger.

We will state the relation between n and m when e is constant in the following special case (treated by Peterson and Birdsall [3], who stated the relation between e and m when n is held constant by the experimenter): we take $p_1 = p_2 = \cdots = p_m = 1/m$ and x a multivariate random variable $x(1), \cdots, x(m)$. Under s_i, suppose that $x(1), \cdots, x(m)$ are independent and that $x(i)$ is normally distributed with mean $\mu > 0$ and unit variance, while the other components of x are normal with zero means and unit variances. Thus there is all-round symmetry. $x(1), \cdots, x(m)$ can be regarded as the inputs on m similar channels. The ith channel is stimulated under s_i. It is readily seen that the optimal procedure is to choose the signal corresponding to the channel with the largest total. It is shown in Appendix 5 that, with this procedure,

$$n\mu^2 = \{1 + [0.64(m-1)^{-1/2} + 0.45]^2\}[\Phi^{-1}(1-e) - \Phi^{-1}(1/m)]^2$$

for those m for which $e < 1 - (1/m)$. Φ^{-1} is the inverse of the normal standardized distribution function. The values of $n\mu^2$ for certain values of e and m have been calculated. If μ is independent of m, then T_d is proportional to $n\mu^2$ and the results are plotted in Figure 1. It can be seen that T_d is very nearly linear against log m, which agrees with some experimental findings in this field.

The question may be raised whether any m-choice task can obey the symmetry condition of the model. Peterson and Birdsall apply the model to the case where an auditory signal is presented in one of four equal periods

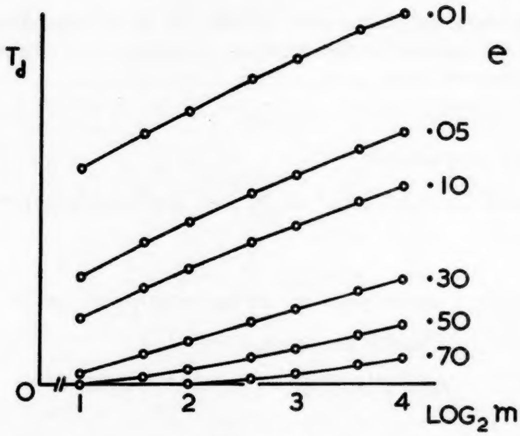


FIGURE 1
The Decision Time (T_d) for Error Rate (e) and Number of Equally Likely Alternatives (m)

of an exposure of S to "white" noise. In this case symmetry is superficially present, but any memory difficulties of S would upset it. We would not expect the model to apply to the case of response to one of m fairly easily discriminable lights arranged in some display, for the noise would be highly positional. However, in the case where the lights are patches of white noise on one of which a low intensity visual signal is superimposed so that response is difficult, the positional effect may not be important and there may be symmetry.

Appendix 1

Let n_{ij} be the sample size for a decision in favor of s_i when s_j is presented. The distribution of n_{ij} is completely determined by its moment generating function, ψ_{ij}. From $A5.1$ of [1], if

$$\phi_i(t) = \sum_x p_i(x)[p_1(x)/p_0(x)]^t,$$

then

$$(1 - \alpha)B^t\psi_{00}[-\log \phi_0(t)] + \alpha A^t\psi_{10}[-\log \phi_0(t)] \equiv 1, \tag{6}$$

$$\beta B^t\psi_{01}[-\log \phi_1(t)] + (1 - \beta)A^t\psi_{11}[-\log \phi_1(t)] \equiv 1, \tag{7}$$

provided the quantities E_i, V_i defined in Appendix 2 are small. If $\alpha < 0.1$ and $\beta < 0.1$ then to a good approximation $A = (1 - \beta)/\alpha$ and $B = \beta/(1 - \alpha)$. Now $\phi_0(1 + u) = \phi_1(u)$; so, putting $t = 1 + u$ in (6) and (7),

$$\beta B^u\psi_{00}[-\log \phi_1(u)] + (1 - \beta)A^u\psi_{10}[-\log \phi_1(u)] \equiv 1,$$

$$(1 - \alpha)B^u\psi_{01}[-\log \phi_0(u)] + \alpha A^u\psi_{11}[-\log \phi_0(u)] \equiv 1.$$

By comparing these equations with (6) and (7), it is found that $\psi_{00} = \psi_{01}$ and $\psi_{10} = \psi_{11}$. Therefore the distributions of n_{00} and n_{01} (and similarly those of n_{10} and n_{11}) are identical.

Appendix 2

In the case of symmetry,

$$\sum_x p_0(x) \log [p_0(x)/p_1(x)] = \sum_x p_1(x) \log [p_1(x)/p_0(x)] = E,$$

and

$$\text{var} \log [p_0(x)/p_1(x)] \text{ under } p_0(x) = \text{var} \log [p_1(x)/p_0(x)] \text{ under } p_1(x) = V.$$

From A:72 of [1], if E and V are small,

$$\bar{n}_0 = J(\alpha, \beta)/E; \quad \bar{n}_1 = J(\beta, \alpha)/E. \tag{8}$$

Therefore

$$\bar{n}_1/\bar{n}_0 = J(\beta, \alpha)/J(\alpha, \beta).$$

By differentiating (6) twice with respect to t and substituting $t = 0$, using (8) and the fact that ψ_{ij} is the moment generating function of n_{ij},

$$v_0 = [VJ(\alpha, \beta)/E^3] - 4[\alpha(1 - \alpha)\bar{n}_1^2/(1 - \alpha - \beta)^2].$$

By symmetry

$$v_1 = [VJ(\beta, \alpha)/E^3] - 4[\beta(1 - \beta)\bar{n}_0^2/(1 - \alpha - \beta)^2].$$

Hence

$$J(\alpha, \beta)v_1 - J(\beta, \alpha)v_0 = 4[J(\beta, \alpha)\alpha(1 - \alpha)\bar{n}_1^2 - J(\alpha, \beta)\beta(1 - \beta)\bar{n}_0^2]/(1 - \alpha - \beta)^2.$$

Appendix 3

If $\alpha < 0.1$ and $\beta < 0.1$ then, by (8), $\bar{T}_d \propto p_0 J(\alpha, \beta) + (1 - p_0)J(\beta, \alpha)$. Keeping e [or $p_0 \alpha + (1 - p_0)\beta$] constant at a value in the range given by $10e < p_0 < 1 - 10e$, the condition on α and β will be satisfied. It is found by the usual methods that the minimum $\bar{T}_d$ is proportional to $J(e, 1 - e) - J(p_0, 1 - p_0)$.

Appendix 4

Let X be the set of all possible values of $x = (x_1, \cdots, x_n)$ and X_i the set of x for which a decision is made for s_i. Then

$$e = \sum_{i=1}^{m} p_i \sum_{x \in X - X_i} p_i(x).$$

Suppose X_i and X_j have a common boundary; then, for e to be a minimum,

it will not be changed by small displacements in this boundary. Hence, on the boundary, $p_i p_i(x) = p_j p_j(x)$; that is, the posterior probability of s_i equals that of s_j. Considering all possible boundaries, the solution is that X_i is the set of x's for which s_i has greater posterior probability than the other signals.

Appendix 5

Write

$$\bar{x}(i) = \sum_{s=1}^{n} x_s(i)/n.$$

Then, under s_1, $\sqrt{n}\bar{x}(1)$ is $N(\sqrt{n}\mu, 1)$ and $\sqrt{n}\bar{x}(i)$ is $N(0, 1)$ for $i \neq 1$. Therefore,

$$\begin{aligned}\alpha_1(\mathfrak{D}) &= \cdots = \alpha_m(\mathfrak{D}) \\ &= 1 - (2\pi)^{-1/2} \int_{-\infty}^{\infty} [\Phi(u)]^{m-1} \exp\left[-\tfrac{1}{2}(u - \sqrt{n}\,\mu)^2\right] du.\end{aligned}$$

On integration by parts,

$$\begin{aligned}(9) \qquad e &= \sum p_i \alpha_i(\mathfrak{D}) \\ &= (m-1)(2\pi)^{-1/2} \int_{-\infty}^{\infty} \Phi(u)^{m-2} \Phi(u - \sqrt{n}\,\mu) \exp\left(-\tfrac{1}{2}u^2\right) du \\ &= e_m(\theta),\end{aligned}$$

say, where $\theta = \sqrt{n}\mu$. Peterson and Birdsall [3] use this form as the basis of their tabulation. However $e_m(\theta) \to 0$ as $\theta \to \infty$ and $e_m(\theta) \to 1$ as $\theta \to -\infty$; while $e_m'(\theta) \leq 0$. Therefore $|e_m'(\theta)|$ is a "probability density function" for θ. The characteristic function and hence the distribution of θ turns out to be the same as that of $v + w$, where $w = \max(v_1, \cdots, v_{m-1})$ and $v, v_1, \cdots, v_{m-1}$ are m independent standard normal variables. Referring to Graph 4.2.2(7) of [4], it can be seen that, for $m < 20$, the first and second moment quotients of w are not very different from those of a normal distribution. Also the addition of v to w will improve normality. Hence θ is approximately normal, agreeing with the calculations of Peterson and Birdsall. If θ is $N(\nu, \sigma^2)$, we determine ν and σ^2 as follows. From (9), $e_m(0) = 1 - (1/m)$. Also $e_m(0) = 1 - \Phi(-\nu/\sigma)$. Therefore

$$\nu/\sigma = -\Phi^{-1}(1/m).$$

Also $\sigma^2 = \text{var } v + \text{var } w$ and from Graph 4.2.2(6) of [4], var $w = [0.64\,(m-1)^{-\frac{1}{2}} + 0.45]^2$ for $m < 20$, which determines σ^2. Putting $e_m(\theta) = e$, the constant error rate,

$$n\mu^2 = \{1 + [0.64(m-1)^{-1/2} + 0.45]^2\}[\Phi^{-1}(1-e) - \Phi^{-1}(1/m)]^2.$$

REFERENCES

[1] Wald, A. *Sequential analysis*, New York: Wiley, 1947.
[2] Peterson, W. W. and Birdsall, T. G. The theory of signal detectability. Tech. Rep. No. 13, Electronic Defense Group, Univ. Michigan, 1953.
[3] Peterson, W. W. and Birdsall, T. G. The probability of a correct decision in a forced choice among M alternatives. Quarterly Prog. Rep. No. 10, Electronic Defense Group, Univ. Michigan, 1954.
[4] Gumbel, E. J. *Statistics of extremes*. New York: Columbia Univ. Press, 1958.

Manuscript received 10/26/59

Revised manuscript received 1/4/60

PSYCHOMETRIKA—VOL. 25, NO. 3
SEPTEMBER, 1960

RELIABILITY FORMULAS FOR INDEPENDENT DECISION DATA WHEN RELIABILITY DATA ARE MATCHED*

NAGESWARI RAJARATNAM

UNIVERSITY OF ILLINOIS

A distinction is made between reliability data and decision data. Each of these sets of data may be matched or independent, depending on whether the same instruments (tests, judges, etc.) are applied to every individual in the group or the instruments to be applied to each individual are selected independently for him. Reliability formulas are developed (for both single observations and for composites of k observations) for the case where reliability data are matched but decision data are independent. Formulas previously reported in the literature are inappropriate for this case.

During the last two decades many writers have approached the problem of reliability through analysis of variance, e.g., Hoyt [5], Ebel [2], Alexander [1], and Harold Webster†. Such an approach not only reveals more clearly the implications of formulas already in the literature, but also leads to new formulas.

This paper deals with the reliability of scores (ratings, etc.) derived from a measuring procedure, i.e., various measuring instruments satisfying a certain description, as distinct from a specific measuring instrument, e.g., Form A of a particular test. The measuring procedure for which a reliability coefficient is sought is defined in terms of a universe of instruments which satisfy a given description. Some examples are: a universe of all possible parallel forms of a particular type of ability test, a universe of all possible judges, a universe of all possible items of a given type, etc. The true score for each individual is defined as the mean score for him on all items, tests, judges, etc., in the universe under consideration.

The reliability coefficient is defined as the ratio of true-score variance to the observed variance expected in any set of data that may be obtained by using a measuring procedure in a specified manner. It indicates how much of the observed variance in a set of data can reasonably be assigned to the variance of true scores.

A reliability coefficient is meant to apply to data which are obtained in order to make some decision about assignments of individuals, about

*This paper was prepared under USPHS Grant M-1839, in which the author is associated with Lee J. Cronbach and Goldine C. Gleser, and is an outgrowth of a more comprehensive conceptualization of reliability being developed in that study. The author is indebted to Dr. Frederic M. Lord for numerous suggestions.

†An unpublished paper prepared in 1958 entitled "A generalization of Kuder-Richardson reliability formula 21."

scientific hypotheses, etc.; hence such data are called decision data. Any sample on which data are (or will be) collected in order to make some decision is called a decision sample. A reliability coefficient is, however, usually estimated from data collected in a special reliability study. Such data will be called reliability data, and the sample on which such data are collected will be referred to as the reliability sample. A reliability coefficient is therefore estimated from data on a reliability sample and is intended to apply to data on a decision sample. While the reliability sample may itself be a decision sample, any one of the many samples to which the measuring procedure is applied in the present or in the future may also serve as a decision sample. All samples (reliability and decision) are considered to be drawn at random from a population in which the investigator is interested. In this paper this population as well as the universe of instruments is assumed to be infinite.

Data are also classified according to the measuring instruments used to generate them. Matched data are obtained by applying the same measuring instrument (or instruments) to every individual in the group. Independent data are obtained by applying to each individual in the group the measuring instrument (or instruments) selected independently for him. Decision data as well as reliability data can be matched or independent. This paper develops formulas for estimating the reliability of independent decision data from matched reliability data. All formulas developed in this paper are new. The only formula of this type in the literature is a special case of one of our formulas developed by Lord [8]. The following are examples of practical problems for which the new formulas are pertinent. (i) Selection decisions are based on ratings of employees by their respective supervisors, while data for the reliability study are obtained by having every member of a group of employees rated by the same supervisors. (ii) Students are selected for special classes on the basis of grade-point averages where each student has taken different courses, but the reliability study involves subjects all of whom have taken the same courses.

The scores for which a reliability coefficient is desired may be single observations or composites of a specified number of single observations. In either case the instruments used to generate data in both the decision sample and the reliability sample are considered to be random samples from the universe of instruments. Where decision data consist of single observations, the investigator has to obtain at least two observations per person in the reliability study. He may then apply to the reliability data formulas developed for single observations. Where decision data are composites, a reliability coefficient may be estimated in two ways. When at least two composites per person are available in the reliability study, the composites may be treated as single observations, the same formulas being used to estimate reliability; or information on the elements which make up the composites may be used to estimate reliability. In the latter case the formulas to be used are based on

an analysis of the elements which make up the composites and are called internal-consistency reliability formulas. If only one composite per person is available, only formulas of the internal-consistency type are applicable. This paper develops formulas for single observations as well as internal-consistency formulas.

Ratios of estimates of true and observed score variance are used to estimate reliability. Following Snedecor [9], Edwards [3], and others, the sample sum of squared deviations divided by degrees of freedom is used to define variance. This gives an unbiased estimate of the population variance as well as estimates of variance for random samples of any size drawn from the same population.

Some writers [4, 6, 7, 8] have defined variance as the sample sum of squared deviations divided by the sample size, and have used this definition to derive reliability formulas. This definition of variance leads to a biased estimate of the population variance and has on that account been discarded by most statisticians. However, formulas based on biased estimates of variances may be used with large samples since the bias is small when samples are large. Although the main argument in this paper will be given in terms of unbiased estimates of population variances, a set of formulas which employ biased estimates will be presented in a separate section in order to show the connection between our formulas and that given by Lord ([8], formula 47).

The argument in the rest of the paper will be stated in terms of ratings by judges, although the same argument may be applied equally well to test and item scores, time samples of behavior, etc.

Notation

X_{pi} = rating of person p by judge i,
M_i = mean for judge i over all persons in the population,
$\bar{X}_{.i}$ = mean rating by judge i of a random sample of persons drawn from the population,
M_p = mean for person p over all judges in the universe,
$\bar{X}_{p.}$ = mean rating of person p by a random sample of judges drawn from the universe,
M = mean for all judges in the universe over all persons in the population,
$\hat{V}$ = an unbiased estimate of a population variance,
$\hat{\rho}$ = ratio of unbiased estimates of population variances of true and observed scores.

Mathematical Model

In the notation given above,

$$X_{pi} = M + (M_p - M) + (M_i - M) + f_{pi}\,, \tag{1}$$

where f_{pi} is a residual component of X_{pi}. In the reliability study k judges

drawn at random from a universe of judges are considered to have rated n persons drawn randomly from the population of persons. Since reliability data are matched each judge rates each of the n individuals.

A two-way analysis of variance of the ratings is as follows.

	Source	df.	Sum of Squares	Mean Square
(2a)	Between persons	$(n-1)$	SS_p	$SS_p/(n-1) = MS_p$
(2b)	Between judges	$(k-1)$	SS_i	$SS_i/(k-1) = MS_i$
(2c)	Residual	$(n-1)(k-1)$	SS_r	$SS_r/(n-1)(k-1) = MS_r$
	Total	$nk-1$	SS_t	

As a consequence of the definition of M_p, M_i, and M,

$$\underset{p}{E}\, f_{pi} = \underset{i}{E}\, f_{pi} = 0.$$

Denote

$$\underset{p}{E}\,(M_p - M)^2, \quad \underset{i}{E}\,(M_i - M)^2, \quad \text{and} \quad \underset{p}{E}\,\underset{i}{E}\, f_{pi}^2 \quad \text{by} \quad V_{M_p}, \quad V_{M_i},$$

and V_f, respectively. Then

$$E(MS_p) = kV_{M_p} + V_f, \tag{3}$$

$$E(MS_i) = nV_{M_i} + V_f, \tag{4}$$

$$E(MS_r) = V_f. \tag{5}$$

The expectations are over all possible samples of judges and for all possible samples of persons. From these equations, unbiased estimates of V_f, V_{M_p}, and V_{M_i} are obtained as follows:

$$\hat{V}_f = MS_r, \tag{6}$$

$$\hat{V}_{M_p} = \frac{1}{k}[MS_p - MS_r], \tag{7}$$

$$\hat{V}_{M_i} = \frac{1}{n}[MS_i - MS_r]. \tag{8}$$

The Reliability of Single Observations

True Variance. Equation (7) above may be used to estimate the variance of true scores.

Observed Variance. Since a reliability coefficient is defined as the ratio of true-score variance to the variance of observed scores in the decision data,

it is necessary to estimate the observed variance expected in the decision data. In independent decision data the observed variance is a function of differences among persons in the three components $(M_p - M)$, $(M_i - M)$, and f_{pi}, cf. (1). If the expected variance of observed scores for the population is denoted by V_X, then

$$V_X = V_{M_p} + V_{M_i} + V_f . \tag{9}$$

Substituting estimates from (6), (7), and (8) in (9),

$$\hat{V}_X = \frac{1}{k}\,\mathrm{MS}_p + \frac{1}{n}\,\mathrm{MS}_i + \left(1 - \frac{1}{k} - \frac{1}{n}\right)\mathrm{MS}_r . \tag{10}$$

Error Variance. The error variance, V_e, may be defined as that part of the observed variance not accounted for by differences in true score. Under this definition, with independent data,

$$\begin{aligned} V_e &= V_X - V_{M_p} \\ &= V_{M_i} + V_f . \end{aligned} \tag{11}$$

Substituting estimates from (6) and (8) in (11),

$$\hat{V}_e = \frac{1}{n}\,\mathrm{MS}_i + \left(\frac{n-1}{n}\right)\mathrm{MS}_r = \mathrm{MS}_{wp} , \tag{12}$$

where

$$\mathrm{MS}_{wp} = \frac{\sum \mathrm{SS}_{wp}}{n(k-1)}, \quad \text{and} \quad \sum \mathrm{SS}_{wp} = \mathrm{SS}_i + \mathrm{SS}_r = \mathrm{SS}_t - \mathrm{SS}_p .$$

Reliability. The formula for estimating reliability is derived from formulas (7) and (10) and is

$$\hat{\rho} = \frac{n(\mathrm{MS}_p - \mathrm{MS}_r)}{n\mathrm{MS}_p + k\mathrm{MS}_i + (nk - n - k)\mathrm{MS}_r}. \tag{13}$$

Dichotomous Case. When ratings or scores are dichotomized and take on the values 0 or 1, versions of the general formulas which are simpler to compute may be obtained. For this purpose let

T_i = number of persons in the reliability sample given a rating of 1 by judge i,

T_p = total score for person p,

T = sum of the k T_i's (= sum of the n T_p's).

Equation (13) then becomes

$$\hat{\rho} = 1 - \frac{(n-1)(kT - \sum_p T_p^2)}{\sum_p T_p^2 + \sum_i T_i^2 - T^2 + (nk - n - k)T}. \tag{14}$$

The Reliability of Composites

Some decisions are based on averages (or sums) of ratings by a specified number of judges (say k). For independent data the set of judges is considered to have been selected randomly and independently for each individual. If reliability data are matched and based on ratings of n persons by k judges as before, the reliability of such composites may be estimated by means of internal-consistency formulas developed in this section. Although the argument here will be stated in terms of means, the identical formulas may be used to estimate the reliability of sums.

The scores for which a reliability coefficient is now desired are the means of k ratings. In the notation of this paper, the mean score for person p, $\bar{X}_{p.}$, may be written as

$$\bar{X}_{p.} = M + (M_p - M) + (\bar{M}_i - M) + \bar{f}_{p.}\,, \tag{15}$$

where $\bar{M}_i$ is the mean of k M_i's and $\bar{f}_{p.}$ is the mean of k f_{pi}'s for person p. In independent data the set of M_i's is different for each individual.

The true score for person p is the expected value of $\bar{X}_{p.}$ over all possible samples of judges from the universe. This is M_p, which is also the true score for single observations. The formula for estimating the variance of true scores for composites is also therefore given by (7).

Since decision data are independent, the expected observed variance is a function of differences among persons in the components $(M_p - M)$, $(\bar{M}_i - M)$, and $\bar{f}_{p.}$, cf. (15). Therefore

$$V_{\bar{X}} = V_{M_p} + \frac{1}{k} V_{M_i} + \frac{1}{k} V_f\,, \tag{16}$$

and

$$\hat{V}_{\bar{X}} = \frac{1}{nk} [n\mathrm{MS}_p + \mathrm{MS}_i - \mathrm{MS}_r]. \tag{17}$$

If error variance is defined as in the previous section, it is estimated by

$$\begin{aligned} \hat{V}_{\bar{e}} &= \frac{1}{k} (\hat{V}_{M_i} + \hat{V}_f) \\ &= \frac{1}{k} \mathrm{MS}_{wp}\,. \end{aligned} \tag{18}$$

The formula for estimating reliability is now

$$\hat{\rho} = \frac{n(\mathrm{MS}_p - \mathrm{MS}_r)}{n\mathrm{MS}_p + \mathrm{MS}_i - \mathrm{MS}_r}. \tag{19}$$

When elements are scored 0 or 1 (19) becomes

$$(20) \qquad \hat{\rho} = 1 - \frac{(n-1)(kT - \sum_p T_p^2)}{(nk - n + 1) \sum_p T_p^2 + k \sum_i T_i^2 - kT^2 - kT}.$$

The principal formulas derived are (13), (14), (19), and (20). All the formulas are new except that an alternate version of formula (20), based on biased estimates of variances, appeared as formula (47) in Lord's recent paper [8].

When decision data are independent and reliability data are matched, the general formulas to be used are (13) to estimate the reliability of single observations and (19) to estimate the reliability of averages (or sums) of k observations. Formulas (14) and (20) are the corresponding formulas for the dichotomous case.

Formulas Based on Biased Estimates

Since the unbiased estimates of variance used in the previous sections are also efficient estimates and since formulas based on unbiased estimates are equally appropriate for use with large as well as small samples, there seems to be no advantage in considering biased estimates. However, one set of biased estimates are considered in this section because they have been used previously by other writers [2, 4, 6, 7, 8] and are on that account of interest. In order to distinguish the estimates used in this section from the unbiased estimates used elsewhere in this paper, the variance estimates considered here will be denoted by $\hat{V}_{(\text{biased})}$, and the reliability estimates by $\hat{\rho}_{(\text{biased})}$.

The following formulas give biased estimates of V_f , V_{M_p} , and V_{M_i} respectively:

$$(6b) \qquad \hat{V}_{f(\text{biased})} = \frac{SS_r}{n(k-1)},$$

$$(7b) \qquad \hat{V}_{M_p(\text{biased})} = \frac{SS_p}{nk} - \frac{SS_r}{nk(k-1)},$$

$$(8b) \qquad \hat{V}_{M_i(\text{biased})} = \frac{SS_i}{n(k-1)}.$$

From these, substituting in (9) for single observations,

$$(10b) \qquad \hat{V}_{X(\text{biased})} = \frac{SS_p}{nk} + \frac{SS_r}{nk} + \frac{SS_i}{n(k-1)}.$$

Now

$$\text{(13b)}\qquad \begin{aligned}\hat{\rho}_{(\text{biased})} &= \frac{\hat{V}_{Mp(\text{biased})}}{\hat{V}_{X(\text{biased})}}\\ &= \frac{\mathrm{SS}_p - \mathrm{SS}_r/(k-1)}{\mathrm{SS}_p + \mathrm{SS}_r + k\mathrm{SS}_i/(k-1)}.\end{aligned}$$

For dichotomously scored data, in the notation given earlier, (13b) becomes

$$\text{(14b)}\qquad \hat{\rho} = 1 - \frac{n(kT - \sum_p T_p^2)}{n(k-1)T - T^2 + \sum_i T_i^2}.$$

When composites of k observations are under consideration,

$$\text{(17b)}\qquad \hat{V}_{\bar{X}(\text{biased})} = \frac{\mathrm{SS}_p}{nk} + \frac{\mathrm{SS}_i}{nk(k-1)},$$

and

$$\text{(19b)}\qquad \hat{\rho}_{(\text{biased})} = \frac{\mathrm{SS}_p - \mathrm{SS}_r/(k-1)}{\mathrm{SS}_p + \mathrm{SS}_i/(k-1)}.$$

For dichotomously scored data (19b) becomes

$$\text{(20b)}\qquad \hat{\rho}_{(\text{biased})} = 1 - \frac{n(kT - \sum_p T_p^2)}{n(k-1)\sum_p T_p^2 - kT^2 + k\sum_i T_i^2}.$$

Equation (20b) may be written

$$\text{(20b}'\text{)}\qquad \hat{\rho}_{(\text{biased})} = \frac{k}{k-1}\left[1 - \frac{\bar{P}\bar{Q} + s_P^2/(k-1)}{s_T^2/k + ks_P^2/(k-1)}\right],$$

where

$$\bar{P} = \frac{T}{nk}, \qquad \bar{Q} = 1 - \bar{P}, \qquad s_P^2 = (k\sum_i T_i^2 - T^2)/n^2k^2,$$

and

$$s_T^2 = (n\sum_p T_p^2 - T^2)/n^2.$$

Formula (20b′) is identical with Lord's formula ([8], formula 47). Lord, however, derived his formula for the purpose of estimating a regression coefficient B, to be used in estimating a person's true score from his observed score on any one of many randomly parallel tests composed of dichotomously scored items. In our notation, Lord's estimation formula is

$$\hat{M}_p = B\bar{X}_{p\cdot} + C, \qquad \text{where } C \text{ is a constant.}$$

Lord's formula treats decision data as independent since characteristics

common to all samples of items drawn from the universe rather than those of one particular form of a test are used to determine B. Reliability data, however, are matched since Lord assumes that every person takes the same form of the test. Since B is a regression coefficient,

$$B = \frac{\sigma_{M_p}\rho_{M_p\bar{X}_{p.}}}{\sigma_{\bar{X}_{p.}}},$$

where ρ and σ are population parameters. But

$$\frac{\sigma_{M_p}}{\sigma_{\bar{X}_p}} = \rho_{M_p\bar{X}_{p.}}, \quad \text{and} \quad \rho^2_{M_p\bar{X}_{p.}} = \rho_{\bar{X}_{p.}\bar{X}_{p.}}.$$

Hence, $B = \rho_{\bar{X}_{p.}\bar{X}_{p.}}$, the expected correlation between two sets of independent data for an infinitely large population. It is therefore not surprising to find that Lord's B is a reliability coefficient of the type considered in this paper.

A Numerical Example

The following is a hypothetical numerical example to illustrate the use of the formulas developed in this paper. Table 1 presents a set of matched reliability data consisting of ratings of 10 persons by 5 judges (i.e., $k = 5$, $n = 10$).

The analysis of variance yields the following information:

Source	df.	Sum of Squares	Mean Square
Between persons	9	93.2	10.36
Between judges	4	20.0	5.0
Residual	36	42.8	1.19
Total	49	156.0	

$$SS_p = 93.2, \qquad MS_p = 10.36,$$

$$SS_i = 20.0, \qquad MS_i = 5.0,$$

$$SS_r = 42.8, \qquad MS_r = 1.19;$$

$$\hat{\rho}_{(13)} = \frac{10\,(10.36 - 1.19)}{103.6 + 25.0 + 35\,(1.19)} = \frac{91.7}{170.25} = .54,$$

$$\hat{\rho}_{(19)} = \frac{91.7}{103.6 + 5.0 - 1.19} = \frac{91.7}{107.41} = .85,$$

$$\hat{\rho}_{(13b)} = \frac{93.2 - (42.8/4)}{93.2 + 42.8 + 5\,(20/4)} = \frac{82.5}{161.0} = .51,$$

TABLE 1

Matched Reliability Data: A Hypothetical Example

		Judges 1	2	3	4	5	$\overline{X}_{p.}$
Persons	1	4	2	5	4	3	3.6
	2	4	4	3	4	4	3.8
	3	5	3	6	5	7	5.2
	4	4	2	3	5	4	3.6
	5	7	7	6	7	9	7.2
	6	6	5	7	6	9	6.6
	7	4	3	3	5	2	3.4
	8	5	5	5	6	7	5.6
	9	7	6	8	4	8	6.6
	10	4	3	4	4	7	4.4
	$\overline{X}_{.i}$	5	4	5	5	6	5.0 $= \overline{X}_{..}$

and

$$\hat{\rho}_{(19b)} = \frac{82.5}{93.2 + (20/4)} = \frac{82.5}{98.2} = .84.$$

Note that $\hat{\rho}_{(13b)} < \hat{\rho}_{(13)}$ and $\hat{\rho}_{(19b)} < \hat{\rho}_{(19)}$, and that

$$\hat{\rho}_{(19)} = \frac{k\hat{\rho}_{(13)}}{1 + (k - 1)\hat{\rho}_{(13)}} \quad \text{and} \quad \hat{\rho}_{(19b)} = \frac{k\hat{\rho}_{(13b)}}{1 + (k - 1)\hat{\rho}_{(13b)}}.$$

REFERENCES

[1] Alexander, H. W. The estimation of reliability when several trials are available. *Psychometrika*, 1947, 12, 79–99.
[2] Ebel, R. L. Estimation of the reliability of ratings. *Psychometrika*, 1951, 16, 407–424.
[3] Edwards, A. L. *Statistical analysis*. (Rev. ed.) New York: Rinehart, 1958.
[4] Horst, P. A generalized expression for the reliability of measures. *Psychometrika*, 1949, 14, 21–31.

[5] Hoyt, C. Test reliability estimated by analysis of variance. *Psychometrika,* 1951, **6,** 153–160.
[6] Kuder, G. F. and Richardson, M. W. The theory of the estimation of test reliability. *Psychometrika,* 1937, **2,** 151–160.
[7] Lord, F. M. Estimating test reliability. *Educ. psychol. Measmt,* 1955, **15,** 325–336.
[8] Lord, M. Statistical inferences about true scores. *Psychometrika,* 1959, **24,** 1–17.
[9] Snedecor, G. W. *Statistical methods.* (5th ed.) Ames, Iowa: Iowa State College Press, 1956.

Manuscript received 7/4/59

Revised manuscript received 11/30/59

PSYCHOMETRIKA—VOL. 25, NO. 3
SEPTEMBER, 1960

A MODEL FOR DETECTION AND RECOGNITION WITH SIGNAL UNCERTAINTY*

ELIZABETH F. SHIPLEY

HARVARD UNIVERSITY†

A model for signal detectability suggested by Luce is extended to situations in which the observer is uncertain of some important characteristic of the signal, such as frequency. By making a single assumption concerning the observer's covert response behavior, two solutions are obtained corresponding to qualitatively different behavior. Decrements in detectability and in recognition with uncertainty are shown to be particular functions of discriminability and detectability of the stimuli in other situations. Relevant experimental data are considered.

It has been found experimentally [7, 9, 10, 11] that an acoustic signal presented in noise is detected more easily when the observer knows the frequency of the signal than when the observer knows only the set of several possible frequencies. Moreover, as the frequency separation of possible signals increases, the decrement in detectability increases [10]. A model of signal detectability based upon decision theory and an assumption resembling Thurstone's discriminal processes has been employed to explain the phenomena using one or the other of the following additional assumptions [1, 10].

(i) The observer functions as a narrow band observer who, at a given time, listens only for signals in a narrow range of frequencies. If two signals are sufficiently separated in frequency, he is unable to detect the one presented when listening for the other. It is further assumed that the observer can change the narrow band of frequencies to which he listens.

(ii) The observer functions as a number of narrow band filters centered upon the several possible signal frequencies. The performance decrement results from the greater amount of noise received through several filters compared to that received through a single filter.

Using the signal detectability model and each of these additional assumptions, it is possible to calculate the loss in detectability. For example, Fig. 1 gives the predictions for a forced-choice experiment in which one of two possible frequencies is employed. In forced-choice experiments several

*This work was supported by grant NSF-G5544 from the National Science Foundation. The author is indebted to R. Duncan Luce for critical discussion of the content and helpful suggestions concerning the form of this paper. Useful comments have been made by David M. Green, Francis W. Irwin, Roger M. Shepard, and W. P. Tanner, Jr.

†Now at University of Pennsylvania.

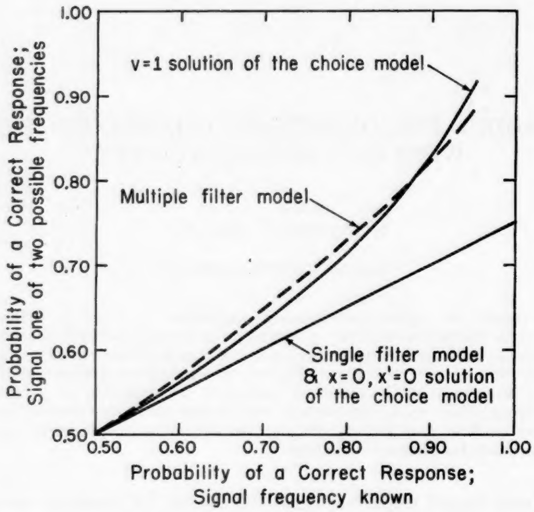


FIGURE 1

Predicted Probability of Correct Response for Three Models of Detection: Forced-Choice Procedure

temporal intervals are defined for the observer; noise is presented in all intervals, but the signal occurs in only one. The observer must report which interval he believes contained the signal. Predictions corresponding to the two assumptions—the single filter model and the multiple filter model—are shown, as are predictions made by the choice model presented here.

This paper offers an alternative explanation of the decrease in detectability in terms of a different model for detection [3]. One feature of the present analysis is that it is not necessary to go outside the basic model in the way that the narrow band filter assumptions do. However, it is necessary to assume that on each trial the observer makes covert recognition responses as well as overt detection responses. Thus, the model also applies to situations in which both recognition and detection responses are made. A second feature of the present work is that it is in no way restricted to acoustic signals or even to any of the usual psycho-physical stimuli; it applies whenever detection and recognition are both possible. However, for concreteness, the analysis is presented in acoustic terminology, with attention confined to two frequencies. The generalization to more frequencies is immediate.

Brief Summary of the Basic Choice Model

We will use two ideas elaborated by Luce [3]. One of these is an assumption that relates response probabilities from overlapping sets of possible

responses (called axiom 1 in [3]). This assumption implies that there exists a positive ratio scale, v, over the set of all possible responses such that the probability of response r being chosen from a subset R of all possible responses is given by the scale value of r, $v(r)$, divided by the sum of the scale values of all the responses in R. Under this assumption, any choice situation that is characterized by the response probabilities is equally characterized by the scale values. Obviously, the ratio scale values for any set of response alternatives may all be multiplied by a constant without changing the probabilities recoverable from the scale values. This property will be used frequently.

The second idea concerns the relations between scale values from a fixed set of response alternatives and two different stimulus conditions, designated 1 and 2. Suppose that the response probabilities in each situation are subject to the basic assumption, and let the corresponding scales be denoted v_1 and v_2. If, for a given response r, $v_2(r)$ is some function of $v_1(r)$, the function depending only upon r and the two stimulus conditions, and if any positive real number is a possible scale value, then it has been argued that the function must be multiplication by a positive constant, i.e.,

$$v_2(r) = \alpha_{1,2,r} v_1(r).$$

For, since the unit of the v scale is unknown, the mathematical form of the function relating the two scales should not presume it to be known. (This is called the independence-of-unit condition in [3].)

Simple Detection and Simple Recognition Situations

In the simplest forced-choice situation, a signal S_1 appears in one of two temporal intervals, and the observer must respond R_1 or R_2 to indicate his judgment that the signal is in the first interval or in the second.

Suppose there is a tendency independent of the signal to select one or the other interval. If only noise is presented, NN, there is a scale value $v_{NN}(R_1)$ for response R_1 and a scale value $v_{NN}(R_2)$ for response R_2. By the first result stated above, when only noise is presented the probability of response R_1 is

$$\frac{v_{NN}(R_1)}{v_{NN}(R_1) + v_{NN}(R_2)}.$$

Setting

$$v = \frac{v_{NN}(R_2)}{v_{NN}(R_1)},$$

then

$$\frac{v_{NN}(R_1)}{v_{NN}(R_1) + v_{NN}(R_2)} = \frac{1}{1 + v}.$$

The quantity v is considered a *response bias* parameter; independent of the signal, it represents the strength of the tendency to respond R_2 when the tendency to respond R_1 is assigned the value unity.

Now consider the stimulus condition S_1N in which the signal S_1 in noise is presented in the first interval and only noise is presented in the second interval. According to the second property mentioned above, if the signal has an effect upon the scale values for the responses R_1 and R_2, then the effect is multiplicative. Let the constant be some number α' for the correct response and α'' for the incorrect response. Thus,

$$v_{S_1N}(R_1) = \alpha' v_{NN}(R_1);$$

$$v_{S_1N}(R_2) = \alpha'' v_{NN}(R_2).$$

Dividing each value by $\alpha'' v_{NN}(R_1)$ and setting $\alpha = \alpha'/\alpha''$,

$$v_{S_1N}(R_1) = \alpha,$$

$$v_{S_1N}(R_2) = v.$$

Next consider the stimulus situation NS_1 in which the signal S_1 appears in the second interval. If the effect of the signal upon the scale value of the correct response is independent of the specific interval containing the signal, then

$$v_{NS_1}(R_1) = \alpha'' v_{NN}(R_1) = 1,$$

$$v_{NS_1}(R_2) = \alpha' v_{NN}(R_2) = \alpha v.$$

In summary, we now have the representation shown in matrix Ia. Ib is the analogue for a signal S_2 which differs from S_1 in frequency or intensity or both.

$$\text{(Ia)} \qquad \begin{array}{c} \\ S_1N \\ NS_1 \end{array}\!\!\begin{array}{c} \begin{array}{cc} R_1 & R_2 \end{array} \\ \begin{bmatrix} \alpha' v_{NN}(R_1) & \alpha'' v_{NN}(R_2) \\ \alpha'' v_{NN}(R_1) & \alpha' v_{NN}(R_2) \end{bmatrix} \end{array} = \begin{array}{c} \\ S_1N \\ NS_1 \end{array}\!\!\begin{array}{c} \begin{array}{cc} R_1 & R_2 \end{array} \\ \begin{bmatrix} \alpha & v \\ 1 & \alpha v \end{bmatrix} \end{array}.$$

$$\text{(Ib)} \qquad \begin{array}{c} \\ S_2N \\ NS_2 \end{array}\!\!\begin{array}{c} \begin{array}{cc} R_1 & R_2 \end{array} \\ \begin{bmatrix} \beta & v \\ 1 & \beta v \end{bmatrix} \end{array}.$$

It should be noted that two matrices of scale values are considered "equal" when they yield the same probability matrix. Thus, for two matrices to be equal, the scale values in corresponding rows must differ only by multiplicative positive constants.

The detection probabilities are immediately recoverable from these representations since the probability of a given response is the scale value of that response divided by the sum of the scale values of all possible responses.

For example, the probability of responding correctly when S_1 is in the first interval is $\alpha/(\alpha + v)$; when S_1 is in the second interval it is $\alpha v/(\alpha v + 1)$. Obviously α and v can be expressed in terms of the response probabilities.

In interpreting such models, we presume the parameters α and β are characteristics of the signal-noise-observer combination but are not under the voluntary control of the observer; they can, however, be affected experimentally by varying signal intensity or frequency, the characteristics of the noise, the time duration of the intervals, the state of the observer (e.g., by drugs), etc. We assume that parameters such as α and β—called *signal parameters* since the model refers to a single observer in a given state—remain unchanged in similar experimental situations, provided that the physical characteristics of the signal and the noise background are the same. This is an empirical assumption that needs to be verified.

On the other hand the response bias v is considered more or less under the control of the observer who can vary it in accord with instructions or when payoffs [6] and perhaps signal intensities are varied. To relate the response bias to payoffs and to properties of the stimuli, an assumption must be made about the observer. One commonly made for this situation is that the observer performs so as to maximize the expected value of his payoff.

A parallel model for the simple recognition situation is equally easy to develop. Either S_1 or S_2 is presented in the noise background, and the observer attempts to identify which one it is; his possible responses, F_1 or F_2, are judgments that frequency 1 or frequency 2 was presented. The model is

$$\text{(II)} \qquad \begin{array}{c} \\ S_1 \\ S_2 \end{array}\!\!\begin{array}{c} \begin{array}{cc} F_1 & F_2 \end{array} \\ \begin{bmatrix} \alpha & \alpha\rho_{12}t \\ \beta\rho_{21} & \beta t \end{bmatrix} \end{array} = \begin{array}{c} \\ S_1 \\ S_2 \end{array}\!\!\begin{array}{c} \begin{array}{cc} F_1 & F_2 \end{array} \\ \begin{bmatrix} 1 & \rho_{12}t \\ \rho_{21} & t \end{bmatrix} \end{array}.$$

Here the response bias is denoted t rather than v, since there is no reason to suppose the recognition bias is the same as the detection bias. The effects of signal S_1 on response F_2 and of signal S_2 on response F_1 have been broken into two multiplicative parts. One part, the signal parameter for detection, α or β, cancels out leaving the signal parameter for recognition, ρ_{ij}. These remaining parameters, ρ_{12} and ρ_{21}, represent the degree of confusion between the two signals.

What are the possible values for these confusion parameters? Clearly, the maximum value of ρ_{ij} is 1; with maximum confusion the probability of responding F_1 is the same whether S_1 or S_2 is presented. Undoubtedly ρ_{ij} is a decreasing function of both the frequency separation and intensity difference of the two signals; the more different the signals, the less the confusion. Furthermore, recognition errors no doubt increase as the intensities of both signals decrease to 0. If neither signal can be detected, how can it be recognized? Now consider two equally detectable signals whose frequency separa-

tion is great enough for minimum confusion. Presumably only the intensities of the two signals determine their confusability. As the signals become more detectable, they are confused less. This suggests taking the minimum of ρ_{ij} to be a function of the reciprocal of the detectability parameter so let us take it to be the simplest function:

$$\rho_{12} = \frac{1}{\alpha}, \qquad \rho_{21} = \frac{1}{\beta}.$$

In addition, the following analysis of detection in the composite forced-choice situation provides an argument for limiting ρ_{ij} to values equal to or greater than the reciprocal of the detectability parameter.

Detection in a Composite Forced-Choice Situation

The recognition and forced-choice procedures may be combined into a more general method in which either one of two frequencies is presented in either one of two intervals and the observer is required both to detect and to recognize the signal. Here both an R and an F response must be made. The model is

$$\text{(III)} \qquad \begin{array}{c} \\ S_1N \\ NS_1 \\ S_2N \\ NS_2 \end{array} \begin{array}{c} \begin{array}{cccc} R_1F_1 & R_1F_2 & R_2F_1 & R_2F_2 \end{array} \\ \begin{bmatrix} \alpha & \alpha\rho_{12}x & y & z \\ 1 & x & \alpha y & \alpha\rho_{12}z \\ \beta\rho_{21} & \beta x & y & z \\ 1 & x & \beta\rho_{21}y & \beta z \end{bmatrix} \end{array}.$$

The Greek symbols have the same meanings as before, and x, y, and z indicate the response bias on R_1F_2, R_2F_1, and R_2F_2, respectively, relative to a bias of unity on R_1F_1.

Next consider the same presentation but with the responses confined to detection. The model for this situation is derived in the same way as that for the simple forced-choice procedure. The response bias parameter v is independent of the signal conditions and two signal detection parameters are employed corresponding to the two signal frequencies. Different symbols are used for signal parameters because the experimental situation is different. However, our analysis will specify a relationship between the signal parameters in the simple and in the composite detection situations. The model for the composite situation is

$$\text{(IV)} \qquad \begin{array}{c} \\ S_1N \\ NS_1 \\ S_2N \\ NS_2 \end{array} \begin{array}{c} \begin{array}{cc} R_1 & R_2 \end{array} \\ \begin{bmatrix} \delta & v \\ 1 & \delta v \\ \gamma & v \\ 1 & \gamma v \end{bmatrix} \end{array}.$$

If we assume that the observer makes a covert recognition response as well as an overt detection response then (III) should yield the probabilities of the detection responses when R_1F_1 and R_1F_2 are combined and R_2F_1 and R_2F_2 are combined. Since (IV) and (III) collapsed on the detection responses yield the same probability matrix, they must be equal. Matrix (V) shows this collapsing of (III).

$$\text{(V)} \qquad \begin{array}{c c} & \begin{array}{cc} \quad R_1 \quad & \quad R_2 \quad \end{array} \\ \begin{array}{c} S_1N \\ NS_1 \\ S_2N \\ NS_2 \end{array} & \begin{bmatrix} \alpha(1 + \rho_{12}x) & y + z \\ 1 + x & \alpha(y + \rho_{12}z) \\ \beta(\rho_{21} + x) & y + z \\ 1 + x & \beta(\rho_{21}y + z) \end{bmatrix} \end{array}.$$

By the definition of equality of scale value matrices, there must be constants c and k such that

$$\delta = c\alpha(1 + \rho_{12}x), \qquad v = c(y + z),$$
$$1 = k(1 + x), \qquad \delta v = k\alpha(y + \rho_{12}z).$$

Hence,

$$\delta/v = \frac{\alpha(1 + \rho_{12}x)}{y + z} \quad \text{and} \quad \delta v = \frac{\alpha(y + \rho_{12}z)}{1 + x}.$$

Similar equations follow from the last two rows. From these two sets of equations,

$$v^2 = \delta v/(\delta/v) = \frac{\alpha(y + \rho_{12}z)(y + z)}{(1 + x)\alpha(1 + \rho_{12}x)};$$
$$v^2 = \gamma v/(\gamma/v) = \frac{\beta(\rho_{21}y + z)(y + z)}{(1 + x)\beta(\rho_{21} + x)}.$$

Equating these expressions and simplifying yields $(z - xy)\,(1 - \rho_{12}\rho_{21}) = 0$, providing neither x, y, nor z is infinite, that is, providing more than one kind of response is made, a condition which seems likely.

We are interested in cases where $\rho_{ij} < 1$. In these cases it must hold that $z = xy$. It follows immediately that $y = v$, i.e., y is the bias parameter for the R_2 response. In the remainder of the analysis y has been replaced by v to emphasize that this bias refers only to the R_2 response relative to the R_1 response.

Multiplying the equations for δv and δ/v and substituting $z = xv$ yields

$$\delta = \frac{\alpha(1 + \rho_{12}x)}{1 + x}.$$

Similarly,

$$\gamma = \frac{\beta(\rho_{21} + x)}{1 + x}.$$

It is clear that $\delta < \alpha$ if and only if $(1 + \rho_{12}x) < (1 + x)$, which is equivalent to $\rho_{12} < 1$. Similarly, $\gamma < \beta$ if and only if $\rho_{21} < 1$. So the signal detection parameters for the composite situation are less than the signal parameters for the simple detection situation if and only if the confusion parameters are less than unity. Also, $\delta \geq 1$, provided $\alpha \geq 1$ and $\alpha\rho_{12} \geq 1$; $\gamma \geq 1$ provided $\beta \geq 1$ and $\beta\rho_{21} \geq 1$.

This last result provides an additional argument for assuming that the minimum value of ρ_{12} is $1/\alpha$ and the minimum value of ρ_{21} is $1/\beta$. If ρ_{12} could be less than $1/\alpha$, then for certain values of α the predicted proportion of *wrong* detection responses would be greater than chance for the composite procedure.

In conclusion, if covert recognition responses occur as well as overt detection responses, and if $\rho_{ij} < 1$, then signal detectability is poorer with two possible frequencies than for signals of a single known frequency. Furthermore, the size of the decrement increases with decreasing ρ_{ij}, which we have argued corresponds to increasing frequency discriminability. Assuming $\alpha = \beta$, $v = 1$, and $x = 1$, the detection probabilities for $\rho_{12} = 1/\alpha$, are shown in Fig. 1 as the $v = 1$ solution of the choice model. Notice that these values are very similar to those predicted by the multiple filter model.

Recognition in a Composite Forced-Choice Situation

We now undertake a parallel analysis in which the roles of recognition and detection are reversed; the observer is required only to identify the frequency when two different signals are used in a forced-choice experiment. The model is

$$\text{(VI)} \qquad \begin{array}{c} \\ S_1N \\ NS_1 \\ S_2N \\ NS_2 \end{array} \begin{array}{c} \begin{array}{cc} F_1 & F_2 \end{array} \\ \begin{bmatrix} 1 & \eta_{12}t \\ 1 & \eta_{12}t \\ \eta_{21} & t \\ \eta_{21} & t \end{bmatrix} \end{array}.$$

As in the simple recognition case, t is the bias on the F_2 response relative to the F_1 response; the confusion parameter η_{12} is employed when S_1 is presented, and η_{21} is employed when S_2 is presented. We take $z = xv$, as a consequence of the analysis of the composite detection situation. Matrix (III) collapsed on the recognition responses is shown as (VIIa). Matrix (VIIb) is equal to (VIIa) with the parameter $x' = 1/x$ introduced.

$$\text{(VIIa)}\qquad \begin{array}{c} \\ S_1N \\ NS_1 \\ S_2N \\ NS_2 \end{array} \begin{array}{c} \begin{array}{cc} F_1 & F_2 \end{array} \\ \begin{bmatrix} \alpha + v & (\alpha\rho_{12} + v)x \\ 1 + \alpha v & (1 + \alpha\rho_{12}v)x \\ \beta\rho_{21} + v & (\beta + v)x \\ 1 + \beta\rho_{21}v & (1 + \beta v)x \end{bmatrix} \end{array}$$

$$\text{(VIIb)}\qquad = \begin{array}{c} \\ S_1N \\ NS_1 \\ S_2N \\ NS_2 \end{array} \begin{array}{c} \begin{array}{cc} F_1 & F_2 \end{array} \\ \begin{bmatrix} (\alpha + v)x' & \alpha\rho_{12} + v \\ (1 + \alpha v)x' & 1 + \alpha\rho_{12}v \\ (\beta\rho_{21} + v)x' & \beta + v \\ (1 + \beta\rho_{21}v)x' & 1 + \beta v \end{bmatrix} \end{array}.$$

Equating (VIIa) and (VI) as was done in the case of detection we obtain from the first two rows the condition

$$\frac{x(\alpha\rho_{12} + v)}{\alpha + v} = \frac{x(1 + \alpha\rho_{12}v)}{1 + \alpha v},$$

which has $x = 0$ as one solution. From (VIIb) and (VI) we obtain from the first two rows

$$\frac{x'(\alpha + v)}{\alpha\rho_{12} + v} = \frac{x'(1 + \alpha v)}{1 + \alpha\rho_{12}v},$$

which has as a solution $x' = 0$. Other solutions to both equations are $v = 1$, $\rho_{12} = 1$, and $\alpha = 0$.

From the last two rows of (VIIa) we get

$$\frac{x(\beta + v)}{\beta\rho_{21} + v} = \frac{x(1 + \beta v)}{1 + \beta\rho_{21}v},$$

and from (VIIb),

$$\frac{x'(\beta\rho_{21} + v)}{\beta + v} = \frac{x'(1 + \beta\rho_{21}v)}{1 + \beta v},$$

which also yield $x = 0$ and $x' = 0$ as solutions as well as $v = 1$, $\rho_{21} = 1$ and $\beta = 0$.

Although our analysis yields several possible solutions, some can be discarded as unlikely general solutions on a priori grounds. To start with the least likely, the $\alpha = 0$ and $\beta = 0$ solutions mean that the probability of a correct detection is zero. If such behavior occurred, the most reasonable interpretations are that the observer failed to understand the instructions or is perverse. Moreover, α and β are considered to be under experimental control.

The solutions $\rho_{ij} = 1$ mean that the two signals are not discriminable. Again, since this is considered to be under experimental control, these are not general solutions. Of course when signals for which $\rho_{ij} = 1$ are used, none of the other solutions need hold.

One likely solution is $v = 1$ which means no bias on the detection responses. Although we know [6] that such a bias can be induced in the simple forced-choice situation by asymmetrical payoffs, perhaps there is no such bias when detection responses are covert. It is an empirical matter to determine if such behavior occurs.

The final possible solutions are $x = 0$ or $x' = 0$ which mean that, independent of the signal, the recognition response is either always F_1 or always F_2. Since the model refers to individual trials, the recognition bias parameter could be $x = 0$ on some trials and $x' = 0$ on the remainder. Such behavior would be obvious empirically in the composite situation when only recognition responses are overt; the proportion of F_1 responses would be the same when S_1 is presented as when S_2 is presented. That is, frequency discrimination apparently would not occur. These solutions may also seem unlikely. Why should the response bias be such as to prevent discrimination, when discrimination is possible, as shown by $\rho_{ij} < 1$ when obtained by some other procedure? In spite of this objection, these solutions appear worthy of consideration because of a result concerning detection responses in the composite situation.

We might expect to find $t = x$ because both parameters appear to refer to the bias on the recognition responses, but unfortunately this cannot be shown because it is not possible to get an expression for t independent of η_{ij}. Nonetheless, in fitting the model to data it seems wise to try $x = t$, in which case,

$$\eta_{12} = \frac{\alpha\rho_{12} + v}{\alpha + v}.$$

We note that $\eta_{12} > \rho_{12}$ if and only if $\rho_{12} < 1$. Thus there will be an increase in the apparent confusion parameter when the observer is uncertain of the interval in which the signal appears.

We have not previously raised the obvious question whether the confusion parameters are symmetrical, i.e., whether $\rho_{ij} = \rho_{ji}$. The very idea of confusion seems symmetrical, but nothing in this analysis leads to it and a desire for economy in our assumptions suggests not making the additional assumption of symmetry. Moreover, certain data [4] suggest $\rho_{ij} \neq \rho_{ji}$. Of course, it is not enough that the confusion matrix be asymmetric, for that can be due to response biases or, when three or more stimuli are used, to large differences in confusion parameters, or to both. But one can proceed as follows. In a large matrix, assume that the response bias is estimated by

the relative frequency of each response, and that for each row i, $\sum_j \rho_{ij}v_j$ is approximately the same, then with the biases known the confusion parameters can be estimated. This was done for Plotkin's data on Morse code signals and it was found, for example, that

$$\frac{\rho_{6B}}{\rho_{B6}} = 2.3.$$

This deviation from symmetry seems severe enough for it to be doubtful that better estimates of the biases and the other confusion parameters could rescue the assumption of symmetry.

Further Considerations of the Detection Problem in the Composite Forced-Choice Situation

So far we have written simple models for detection and for recognition in the composite forced-choice situations and then assumed that the observer always makes both a detection and a recognition response in this situation. By collapsing the matrix containing scale values for four responses on the recognition responses we found two likely general solutions. Now consider these solutions when only overt detection responses are made in the composite situation. The first solution, $v = 1$, where v is the R_2 response bias, does not affect the values of the apparent signal parameters δ and γ. The second solution says that the observer either sets $x = 0$ or sets $x' = 0$ ($x' = 1/x$); if this solution holds it seems likely that sometimes $x = 0$ and other times $x' = 0$. Matrix (III) becomes (VIII) when $x = 0$ (with $y = v$).

$$\text{(VIII)} \qquad \begin{array}{c} \\ S_1N \\ NS_1 \\ S_2N \\ NS_2 \end{array} \begin{array}{c} \begin{array}{cc} R_1 & R_2 \end{array} \\ \begin{bmatrix} \alpha & v \\ 1 & \alpha v \\ \beta\rho_{21} & v \\ 1 & \beta\rho_{21}v \end{bmatrix} \end{array}.$$

When $x' = 1/x = 0$, (III) becomes (IX) (with $z = xv$).

$$\text{(IX)} \qquad \begin{array}{c} \\ S_1N \\ NS_1 \\ S_2N \\ NS_2 \end{array} \begin{array}{c} \begin{array}{cc} R_1 & R_2 \end{array} \\ \begin{bmatrix} \alpha\rho_{12} & v \\ 1 & \alpha\rho_{12}v \\ \beta & v \\ 1 & \beta v \end{bmatrix} \end{array}.$$

On a given trial if we let p equal the probability that $x = 0$, and $1 - p$

equal the probability that $x' = 0$, then, for example, the probability of a correct detection given stimulus S_1N is

$$P_{S_1N}(R_1) = p\,\frac{\alpha}{\alpha + v} + (1 - p)\,\frac{\alpha\rho_{12}}{\alpha\rho_{12} + v}.$$

The term $\alpha/(\alpha + v)$ is simply the probability of a correct response when the observer knows the signal frequency (Ia). To examine the second term, suppose the frequencies are sufficiently different so that ρ_{12} is near its minimum value, which we have previously argued may be $1/\alpha$. Then the probability of an R_1 response is $1/(1 + v)$ when $x' = 0$, which means that it depends only upon the response bias, not on the signal intensity. These interpretations of the individual terms cannot be checked because it is not possible to distinguish the $x = 0$ trials from the $x' = 0$ trials when only detection responses are made. However, by assuming some value for p, the over-all probability of a correct detection can be estimated and compared with data. The values calculated are the same as those for the single filter assumption (Fig. 1). This is no accident. In that model it is assumed that the observer (i) listens with probability p for S_1, in which case the probability of a correct response when S_1N is presented is the same as in the simple forced-choice situation—call it P_{S_1N}, and (ii) listens with probability $1 - p$ for S_2, in which case he chooses R_1 with some probability Q. Thus,

$$P_{S_1N}(R_1) = pP_{S_1N} + (1 - p)Q.$$

Setting $Q = 1/(1 + v)$ gives the same results as the $x = 0$ or $x' = 0$ choice model when $\rho_{12} = 1/\alpha$. So we have two likely solutions from the analysis, one of which imposes restrictions on the bias on the detection responses and the other imposes restrictions on the bias on the recognition responses. These solutions correspond, in terms of predicted detection behavior in the composite situation, to the multiple filter and single filter signal detectability models respectively.

Detection and Recognition in a Composite Yes-No Situation

In the yes-no situation a single time interval is used, sometimes this interval contains a signal and sometimes it does not. This can be analyzed in much the same way as the forced-choice situation. However, to obtain comparable results it must be assumed in (XI) that $s = qr$; i.e., that the recognition response bias (the bias on F_1 or F_2) is independent of the detection response bias (the bias on Y or N) in the composite situation. This is directly analogous to the result $z = xy$ in (III).

Because of the similarity of the analysis, only the matrices and the conclusions are presented.

$$\text{(Xa, b)} \qquad \begin{array}{c} \\ S_1 \\ N \end{array}\!\!\begin{array}{c} \begin{array}{cc} Y & N \end{array} \\ \begin{bmatrix} \alpha & u \\ 1 & u \end{bmatrix} \end{array} \quad \begin{array}{c} \\ S_2 \\ N \end{array}\!\!\begin{array}{c} \begin{array}{cc} Y & N \end{array} \\ \begin{bmatrix} \beta & u \\ 1 & u \end{bmatrix} \end{array}.$$

$$\text{(XI)} \qquad \begin{array}{c} \\ S_1 \\ S_2 \\ N \end{array}\!\!\begin{array}{c} \begin{array}{cccc} YF_1 & YF_2 & NF_1 & NF_2 \end{array} \\ \begin{bmatrix} \alpha & \alpha\rho_{12}q & r & s \\ \beta\rho_{21} & \beta q & r & s \\ 1 & q & r & s \end{bmatrix} \end{array} \quad \begin{array}{c} \\ S_1 \\ S_2 \\ N \end{array}\!\!\begin{array}{c} \begin{array}{cccc} YF_1 & YF_2 & NF_1 & NF_2 \end{array} \\ \begin{bmatrix} \alpha & \alpha\rho_{12}q & r & qr \\ \beta\rho_{21} & \beta q & r & qr \\ 1 & q & r & qr \end{bmatrix} \end{array}.$$

$$\text{(XII)} \qquad \begin{array}{c} \\ S_1 \\ S_2 \\ N \end{array}\!\!\begin{array}{c} \begin{array}{cc} Y & N \end{array} \\ \begin{bmatrix} \delta & u \\ \gamma & u \\ 1 & u \end{bmatrix} \end{array} = \begin{array}{c} \\ S_1 \\ S_2 \\ N \end{array}\!\!\begin{array}{c} \begin{array}{cc} Y & N \end{array} \\ \begin{bmatrix} \alpha(1 + \rho_{12}q) & r(1 + q) \\ \beta(\rho_{21} + q) & r(1 + q) \\ 1 + q & r(1 + q) \end{bmatrix} \end{array}.$$

$$\text{(XIII)} \qquad \begin{array}{c} \\ S_1 \\ S_2 \\ N \end{array}\!\!\begin{array}{c} \begin{array}{cc} F_1 & F_2 \end{array} \\ \begin{bmatrix} 1 & \eta_{12}t \\ \eta_{21} & t \\ 1 & t \end{bmatrix} \end{array} = \begin{array}{c} \\ S_1 \\ S_2 \\ N \end{array}\!\!\begin{array}{c} \begin{array}{cc} F_1 & F_2 \end{array} \\ \begin{bmatrix} \alpha + r & q(\alpha\rho_{12} + r) \\ \beta\rho_{21} + r & q(\beta + r) \\ 1 + r & q(1 + r) \end{bmatrix} \end{array}.$$

It follows immediately from (XII) that $u = r$ and from (XIII) that $t = q$. Further, from (XII),

$$\delta = \frac{\alpha(1 + \rho_{12}t)}{1 + t},$$

$$\gamma = \frac{\beta(\rho_{21} + t)}{1 + t}.$$

As before, $\delta < \alpha$ if and only if $\rho_{12} < 1$,
$\gamma < \beta$ if and only if $\rho_{21} < 1$.

We predict a reduction in effective detectability when the signal can be one of two possible frequencies. And again the reduction in detectability is greater, the greater the discriminability of the signals. Moreover, if the bias parameter for the recognition responses is the same in the forced-choice and yes-no situations, i.e., $x = t$, then both situations yield the same apparent signal detectability parameter. However, unlike the forced-choice situation, in the yes-no situation only one explanation for a decrement in detection follows from the choice model.

From (XIII) it follows that

$$\eta_{12} = \frac{\alpha\rho_{12} + u}{\alpha + u},$$

$$\eta_{21} = \frac{\beta\rho_{21} + u}{\beta + u}.$$

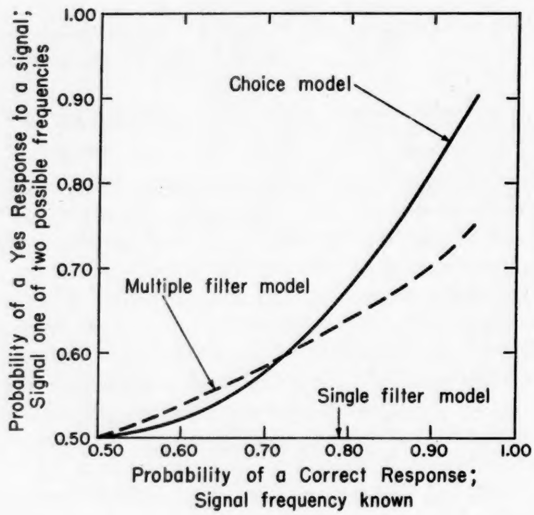


FIGURE 2

Predicted Probability of Correct Response for Three Models of Detection: Yes-No Procedure

We observe that $1 > \eta_{12} > \rho_{12}$ if and only if $\rho_{12} < 1$,
$1 > \eta_{21} > \rho_{21}$ if and only if $\rho_{21} < 1$.

Thus, the effective confusion parameter is larger in the composite situation provided that some discrimination of frequencies occurs in the simple recognition situation. Moreover, the increase depends upon the bias of the detection responses in the same way as in the forced-choice situation.

Examples of predicted detection probabilities for the composite yes-no procedure are shown in Fig. 2, both for the present choice model and for the two narrow band filter models. In order to facilitate comparison of the models, values for the probability of a *yes* response when a signal is given were calculated assuming the same false alarm rate, $P_N(Y)$, in the simple and in the composite situation. Only for relatively high signal intensities do the choice model and the multiple filter model give sizably different predictions.

Some Relevant Experimental Evidence

In some studies [9, 10] the single filter assumption and in another study [11] the multiple filter assumption have been reported to provide a better fit to the data. In one study [7] the single filter assumption made the better

prediction for one observer and the multiple filter assumption for the others. In view of the correspondence between the predictions of the two solutions of the choice model and of the two filter assumptions, there is good reason for believing that both solutions of the choice model hold. It should be noted that all of the above studies employed a forced-choice procedure; in our analysis two solutions appear only for the forced-choice procedure.

There is evidence that the decrement in detection when the observer is uncertain of the signal frequency is greater the greater the frequency separation of the signals [10]. Moreover, there is some evidence that the decrement is greater for four possible frequencies than for two possible frequencies [11]. Also the decrement is greater for signals of shorter duration [9, 10].

The first two findings are certainly consistent with the analysis presented here. We have shown that detection should be poorer for signals which are better discriminated in a recognition procedure, and the data indicate that discriminability increases with frequency separation [8]. It can be readily shown that the apparent signal parameter decreases as the number of signal frequencies increases, providing the confusion parameters are of the same magnitude.

The finding that the decrement in detection is greater for signals of briefer duration has been considered evidence supporting the single filter assumption [8]. The argument is that when the observer has sufficient time he can listen successively to each of the possible frequencies; with very brief signals he can listen only to a single frequency. It cannot be known whether these data are at variance with the present model until recognition data are obtained for signals of the durations and intensities employed. It should be noted that the briefer signals were of higher intensity in these studies so that it is not inconceivable that recognition is better for these signals which, according to the choice model, leads to lower detectability with uncertainty.

Although the available data are encouraging, it seems clear that for an adequate test of the present model data must be collected under a variety of conditions for each observer. Specifically, the experimental conditions described by matrices (Ia), (Ib), (II), (III), (IV), (VI), (Xa), (Xb), (XI), (XII), and (XIII) should be used. Such a study is underway.

Discussion

This analysis of the detection and recognition problem has some features that merit repeating. Working within the framework of a choice model, it has been possible to establish not only that the detectability of one of two possible signals must in general be poorer than that of one, but also to state how the decrement depends upon an experimental measure of the discriminability of the signals. The main assumption was that the observer made a recognition response covertly when overtly he made only a detection

response and vice versa. This assumption refers only to the observer's responses. No extra-model assumptions concerning the hearing mechanism, such as the narrow band filter assumptions previously employed, are needed to account for the decrement.

The assumption on which the present model is based becomes more plausible in view of the results of experiments by Lawrence and Coles [2] and by Pollack [5] on restricted response classes. In the latter study, the set of possible stimulus and response alternatives in a word recognition task was restricted either before the stimulus was presented or after the subject observed the stimulus, but before he responded. The probability of a correct response was apparently independent of the number of response alternatives given prior to the observation. It depended strongly upon the number of alternatives available immediately prior to the response. This suggests that uncertainty with respect to the signal in a detection situation affects only the response mechanism and not the observing mechanism. Obviously the present model does not predict a decrement in detection if the observer is told the signal frequency following each observation but prior to his response.

The formal structure of the model for the composite forced-choice detection procedure led to two different solutions. One solution yields predictions which are almost the same as those of a multiple filter model of signal detectability. The other solution of the present model yields predictions which are identical to those of a single filter model. Thus, where the present model can account for qualitative differences in behavior with a single set of assumptions, the narrow band filter models require a different assumption for each of two kinds of behavior.

Finally, the generality of the present model should be mentioned again. It applies wherever the basic choice axiom holds and both detection and recognition responses are possible. With auditory stimuli such as words or with complex visual stimuli, it is difficult to see how to apply the concept of the narrow band filter mechanisms.

REFERENCES

[1] Green, D. M. Detection of multiple component signals in noise. *J. acoust. Soc. Amer.*, 1958, **30**, 904–911.

[2] Lawrence, D. H. and Coles, G. R. Accuracy of recognition with alternatives before and after the stimulus. *J. exp. Psychol.*, 1954, **47**, 208–214.

[3] Luce, R. D. *Individual choice behavior: a theoretical analysis.* New York: Wiley, 1959.

[4] Plotkin, L. Stimulus generalization in Morse code learning. *Arch. Psychol.*, 1943, 287, 5–39.

[5] Pollack, I. Message-uncertainty and message-reception. *J. acoust. Soc. Amer.*, 1959, **31**, 1500–1508.

[6] Swets, J. A. and Birdsall, T. G. The human use of information, III. Decision-making in signal detection and recognition situations involving multiple alternatives. *Trans. I. R. E. Professional Group on Information Theory*, 1956, 138–165.

[7] Swets, J. A., Shipley, E. F., McKey, M. J., and Green, D. M. Multiple observations of signals in noise. *J. acoust. Soc. Amer.*, 1959, **31**, 514–521.

[8] Tanner, W. P., Jr. Theory of recognition. *J. acoust. Soc. Amer.*, 1956, **28**, 882–888.

[9] Tanner, W. P., Jr. and Norman, R. Z. The human use of information, II. Signal detection for the case of an unknown signal parameter. *Trans. I. R. E. Professional Group on Information Theory*, 1954, 222–226.

[10] Tanner, W. P., Jr., Swets, J. A., and Green, D. M. Some general properties of the hearing mechanism. Tech. Rep. No. 30, Electronic Defense Group, Univ. Michigan, 1956.

[11] Veniar, F. A. Signal detection as a function of frequency ensemble, I. *J. acoust. Soc. Amer.*, 1958, **30**, 1020–1024.

Manuscript received 5/16/59

Revised manuscript received 1/20/60

PSYCHOMETRIKA—VOL. 25, NO. 3
SEPTEMBER, 1960

A PROPOSED VARIATION OF THE MATCHING TECHNIQUE

GEORGE H. WEINBERG, FRITZ A. FLUCKIGER, AND CLARENCE A. TRIPP
THE HANDWRITING INSTITUTE, NEW YORK CITY

A matching procedure is proposed by which a judge may use the same items in different matches. The K items in one group most likely to include the correct match with each item in the other are selected. Inclusion of the correct match among the K items chosen is defined as a success. The distribution of the number of successes is discussed. Tables are presented showing the number of successes needed for significance for various values of K and of N, the number of items in each group.

The usual matching technique consists of giving a subject two sets of N items such that each item in one set corresponds in some way to an item in the other. The items in each set are presented in randomized order and the subject is asked to make the proper pairings. This technique has been used often in connection with projective techniques. For instance, in some studies judges have been asked to match personality descriptions of individuals with their handwritings [1, 4, 7].

The chief advantage of the matching technique is that it permits a test of whether a judge can make proper identifications without forcing him to identify the specific cues he is using. The number of correct pairings is the statistic used in the attempt to reject the hypothesis that the pairings are being made in random fashion. The distribution of the number of correct pairings under this null hypothesis has been discussed by Chapman [2, 3].

The matching technique, as ordinarily used, has serious drawbacks, most of which stem from the fact that success or failure on any one matching influences the probability of success or failure on the others. Vernon [6] and Secord [5] among others, have discussed some of these limitations. For instance, they noted that since elimination is used in making the matchings, the order of presentation of the items is important. The fact that one mistake necessitates another means that the easiest order is the one in which the simplest matchings are done first. The effect of the demand that each item appear in only one matching is to decrease the power of the experiment in the same way that reducing a sample size reduces power in the usual psychological experiment. In essence, the first few matchings become highly determinative.

A second and perhaps more important defect of the matching method is that it demands an unnecessarily high degree of precision from the judge. For instance, suppose it is claimed that the handwriting samples of individuals

TABLE 1

Number of Successes Needed for Significance at the .05 Level for Various Values of N and K

N	K									
	1	2	3	4	5	6	7	8	9	10
7	4	5	6	7	*					
8	4	5	6	7	8	*				
9	4	5	6	7	8	9	*			
10	4	5	6	7	8	9	10	*		
11	4	5	6	8	9	10	10	11	*	
12	4	5	6	8	9	10	11	11	12	*
13	4	5	7	8	9	10	11	12	13	13
14	4	5	7	8	9	10	11	12	13	14
15	4	5	7	8	9	10	11	12	13	14
16	4	5	7	8	9	10	11	12	13	14
17	4	5	7	8	9	10	11	12	13	14
18	4	5	7	8	9	10	11	12	13	14
19	4	5	7	8	9	10	11	12	13	15
20	4	5	7	8	9	10	11	13	14	15
21	4	5	7	8	9	10	11	13	14	15
22	4	5	7	8	9	10	11	13	14	15
23	4	5	7	8	9	10	12	13	14	15
24	4	5	7	8	9	10	12	13	14	15
25	4	5	7	8	9	11	12	13	14	15
30	4	5	7	8	9	11	12	13	14	15
35	4	5	7	8	10	11	12	13	14	15
40	4	5	7	8	10	11	12	13	14	15
45	4	5	7	8	10	11	12	13	14	16

* Significance at the .05 level cannot be obtained when $K \geqq$ than the given value.

in some way tend to reflect their personalities. A matching study is designed in which N writing samples and N brief personality sketches of the individuals who wrote them are collected. The judge who is asked to match the sketches of the individuals with the handwriting samples usually starts by associating some small number of personality sketches (say K of them) with a particular sample. His tentative conclusion, if verbalized at this point, would be that any one of these K people selected from N might have produced the particular handwriting sample.

The constraints of the usual matching method now force the judge to choose one of these K people and to associate that person with the particular sample. In essence, the data are used to choose a subset from among the sketches and then the final choice is made almost at random. Typically, the judge reports great discomfort when forced to make this final choice, for he knows that at this point he is guessing.

In contrast, the judge may be allowed to associate some fixed number of sketches with each handwriting sample. The inclusion of the one correct

TABLE 2

Number of Successes Needed for Significance at the .01 Level for Various Values of *N* and K

N	K 1	2	3	4	5	6	7	8	9	10
7	4	6	7	*						
8	4	6	7	8	*					
9	4	6	7	8	9	*				
10	4	6	7	8	9	10	*			
11	4	6	7	8	10	11	11	*		
12	4	6	7	9	10	11	12	12	*	
13	5	6	8	9	10	11	12	13	13	*
14	5	6	8	9	10	11	12	13	14	14
15	5	6	8	9	10	11	12	13	14	15
16	5	6	8	9	10	11	12	13	14	15
17	5	6	8	9	10	11	12	13	14	15
18	5	6	8	9	10	11	13	14	15	15
19	5	6	8	9	10	12	13	14	15	16
20	5	6	8	9	10	12	13	14	15	16
21	5	6	8	9	10	12	13	14	15	16
22	5	6	8	9	11	12	13	14	15	16
23	5	6	8	9	11	12	13	14	15	16
24	5	6	8	9	11	12	13	14	15	16
25	5	6	8	9	11	12	13	14	15	16
30	5	6	8	9	11	12	13	14	16	17
35	5	6	8	9	11	12	13	15	16	17
40	5	6	8	10	11	12	14	15	16	17
45	5	6	8	10	11	12	14	15	16	17

* Significance at the .01 level cannot be obtained when K $\geq$ than the given value.

match in the subset of selected sketches may be counted as a success. An error is made when the selected sketches do not include the correct match. The purpose is to demonstrate a relationship between individuals and their handwriting productions, and not to demonstrate an isomorphism between individuals and their handwritings.

Consider the situation in which the judge is instructed to match exactly K sketches with each sample. The instructions are to pick the K individuals most likely to have produced each sample. It should be emphasized that a personality sketch may be used any number of times. Once again the number of successes is a random variable which can be used to test whether the judge is associating the sketches with the handwriting samples in a random way.

To be specific, say that there are 20 samples and 20 sketches ($N = 20$). The judge is told to associate with each handwriting sample the names of the three people most likely to have produced it ($K = 3$). Now the order of presentation does not matter. Under the hypothesis that all the associations are being made by chance, the number of successes has a binomial distri-

TABLE 3

Number of Successes Needed for Significance at the .001 Level
for Various Values of N and K

N	K									
	1	2	3	4	5	6	7	8	9	10
7	5	7	*							
8	5	7	8	*						
9	5	7	8	9	*					
10	6	7	9	10	10	*				
11	6	7	9	10	11	11	*			
12	6	7	9	10	11	12	*			
13	6	7	9	10	11	12	13	*		
14	6	7	9	10	11	12	13	14	*	
15	6	8	9	11	12	13	14	15	15	*
16	6	8	9	11	12	13	14	15	16	16
17	6	8	9	11	12	13	14	15	16	16
18	6	8	9	11	12	13	14	15	16	17
19	6	8	9	11	12	13	14	15	16	17
20	6	8	9	11	12	13	15	16	17	17
21	6	8	9	11	12	14	15	16	17	18
22	6	8	9	11	12	14	15	16	17	18
23	6	8	10	11	12	14	15	16	17	18
24	6	8	10	11	12	14	15	16	17	18
25	6	8	10	11	13	14	15	16	17	18
30	6	8	10	11	13	14	15	17	18	19
35	6	8	10	11	13	14	16	17	18	19
40	6	8	10	11	13	14	16	17	18	20
45	6	8	10	12	13	15	16	17	19	20

* Significance at the .001 level cannot be obtained when K ≥ than the given value.

bution. The probability of a success is K/N. The probability of exactly v successes is

(1) $$\frac{N!}{(N-V)!\,V!}\left[\frac{K}{N}\right]^{V}\left[\frac{N-K}{N}\right]^{N-V}.$$

As N increases, the number of successes needed for significance for each particular K reaches an asymptotic limit. The probability of v successes approaches

(2) $$e^{-K}\frac{K^{V}}{V!},$$

which is the Poisson form. The probability of more than v successes is

(3) $$1 - e^{-K}\sum_{S=0}^{V}(K^{S}/S!).$$

Thus, it turns out that when $K = 1$, four successes are needed for significance at the .05 level for any $N \geq 7$. When $K = 3$, seven successes are needed for significance at the .05 level so long as $N \geq 13$.

Note that when $K = 1$, each match becomes the pairing of a single

sample from one group with a sample from the other. However, the fact that the judge may use the same sample for various matches gives him a freedom which he does not have with the ordinary matching method.

Tables 1–3 indicate the number of successes needed for significance at the .05, .01, and .001 levels for various values of N and K. Suppose, for example, that a judge is given the handwriting samples and personality sketches of 19 people ($N = 19$). He associates three sketches with each sample ($K = 3$). According to Table 1, seven or more successes would lead us to reject the hypothesis that the judge is associating the personality sketches with the handwriting in a random way.

The proposed method is preferable to the usual method of carrying out a matching study for two reasons. First, the probability of a success is not a function of previous successes or failures. Thus many defects of the matching method, noted by its critics, are eliminated. In the second place, the proposed method is an attempt to establish evidence in favor of the thesis which is nearly always implicit, that the items in one category provide meaningful information about the items in the other category.

REFERENCES

[1] Allport, G. W. and Vernon, P. E. *Studies in expressive movement.* New York: Macmillan, 1933.

[2] Chapman, D. W. The statistics of the method of correct matchings. *Amer. J. Psychol.*, 1934, **46**, 287–298.

[3] Chapman, D. W. The generalized problem of correct matchings. *Ann. math. Statist.*, 1935, **6**, 85–95.

[4] Eysenck, H. J. Graphological analysis and psychiatry: an experimental study. *Brit. J. Psychol.*, 1945, **35**, 70–81.

[5] Secord, P. F. A note on the problem of homogeneity-heterogeneity in the use of the matching method in personality studies. *Psychol. Bull.*, 1952, **49**, 41–42.

[6] Vernon, P. E. The matching method applied to investigations of personality. *Psychol. Bull.*, 1936, **33**, 149–177.

[7] Wells, F. L. Personal history, handwriting, and specific behavior. *Charact. & Pers.*, 1946, **14**, 295–314.

Manuscript received 7/24/59

Revised manuscript received 11/10/59

PSYCHOMETRIKA—VOL. 25, NO. 3
SEPTEMBER, 1960

ON THE WILSON TESTS*

N. DONALD YLVISAKER

COLUMBIA UNIVERSITY

A general critical analysis of the median tests proposed by Wilson for certain analysis of variance hypotheses is presented. Specifically, discrepancies between the purported and actual approximate distributions of some of the test statistics are noted. Validity and power of the resulting tests are discussed.

Wilson [4] has proposed a set of median tests for certain analysis of variance hypotheses. These tests have received some critical attention in the literature, [e.g., 1, 2]. It is the purpose of this paper to present, as descriptively as possible, a general critical analysis of these tests. This analysis will be done exclusively for the two-way design problem although Wilson has proposed generalizations of his test statistics for n-way designs. The difficulties raised here concerning the former problem will also apply to these generalizations.

The Wilson Tests

The two classifications are labeled row and column and are indexed $i = 1, \cdots, r$, and $j = 1, \cdots, c$, respectively. With this labeling, there is a natural correspondence established between the classifications and the cells of a two-way table.

To the classification (i, j) let there correspond a random variable X_{ij} with distribution function F_{ij} ; further suppose that F_{ij} differs from $F_{i'j'}$ only in location. From this last assumption it follows that the F_{ij} are of the form

$$F_{ij}(x) = F(x - \nu_{ij}),$$

for all x and some choice of the distribution function F. Here the ν_{ij} represent location parameters, corresponding to the choice of F; they may be written uniquely in the form

$$\nu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} ,$$

*This work was sponsored in part by the Office of Naval Research while the author was at Stanford University. Reproduction in whole or in part is permitted for any purpose of the United States Government. The author wishes to thank Professors Fred C. Andrews, Lincoln E. Moses, and David L. Wallace for their helpful criticisms and suggestions in the writing of this paper.

subject to the conditions

$$\sum_i \alpha_i = \sum_j \beta_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0 \quad (i = 1, \cdots, r;\ j = 1, \cdots, c).$$

The α_i, β_j, and γ_{ij} represent row, column, and interaction effects, respectively. The problem under consideration is the following: given n_{ij} independent observations on X_{ij}, test for the presence of row, column, or interaction effects.

In order to describe the proposed tests, the following notation is necessary.

$n \;= \sum n_{ij}$, the total number of observations,
$M \;=$ the median of all n observations,
$bf_{ij} \;=$ the number of observations on X_{ij} which are less than M, i.e., below the over-all median, with
$n_b \;= \sum bf_{ij}$,
$\chi^2_q(\alpha) =$ the upper α-point of the tabulated χ^2 distribution with q degrees of freedom.

Thus M and the bf_{ij} are random variables which are determined by the observations on the X_{ij}. The test statistics together with their corresponding hypotheses are as follows.

1. Test the hypothesis $H_T : \alpha_i = \beta_j = \gamma_{ij} = 0$ (for all i and j), the hypothesis of homogeneity, by rejecting if $M_T \geq \chi^2_{rc-1}(\alpha)$, where

$$M_T = \sum_{i,j} \left[\frac{[bf_{ij} - n_{ij}(n_b/n)]^2}{n_{ij}\cdot(n_b/n)} + \frac{[n_{ij} - bf_{ij} - n_{ij}(n - n_b)/n]^2}{n_{ij}\cdot(n - n_b)/n} \right].$$

2. Test the hypothesis $H_R : \alpha_i = \gamma_{ij} = 0$ (for all i and j), the hypothesis of no row effects, by rejecting if $M_R \geq \chi^2_{r-1}(\alpha)$, where

$$M_R = \sum_i \left[\frac{[bf_{i.} - n_{i.}(n_b/n)]^2}{n_{i.}(n_b/n)} + \frac{[n_{i.} - bf_{i.} - n_{i.}(n - n_b)/n]^2}{n_{i.}(n - n_b)/n} \right],$$

with the dot notation indicating the sum over the corresponding subscript.

3. Test the hypothesis $H_C : \beta_j = \gamma_{ij} = 0$ (for all i and j), the hypothesis of no column effects, by rejecting if $M_C \geq \chi^2_{c-1}(\alpha)$, where M_C is the column analogue of M_R.

4. Test the hypothesis $H_I : \gamma_{ij} = 0$ (for all i and j), the hypothesis of no interaction, by rejecting if $M_I \geq \chi^2_{(r-1)(c-1)}(\alpha)$, where $M_I = M_T - M_R - M_C$.

These tests are proposed for large samples in the sense that is usual for χ^2 tests on counted data, i.e., that the test statistics have approximate χ^2

distributions under the associated null hypotheses provided the sample sizes are sufficiently large.

Distribution Theory under the Null Hypotheses

It can be shown by algebraic manipulation that M_T is, aside from a factor of $(n - 1)/n$, the Mood-Brown statistic (cf. [3], p. 398). It then follows that M_T is distributed approximately as χ^2_{rc-1} under H_T [3].

In order to make the discussion of the distribution of M_R meaningful, some notion of approximating distributions must be introduced. Suppose the distribution of M_R , under the hypothesis H_R , is studied as the sample sizes become large in fixed ratios, i.e., suppose $n_{ij}(t) = s_{ij}t$, where the s_{ij} represent weighting factors and where the parameter t is allowed to grow large. Then it must be expected that the distribution function of $M_R(t)$ (the test statistic here depends on t) should be close to the distribution function of a χ^2_{r-1} random variable provided t is (equivalently, the sample sizes are) sufficiently large. This analysis has been carried out in [5] and two distinct results will be stated and illustrated here.

(a) If H_R is *true*, $M_R(t)$ will in general tend to infinity in probability as $t \rightarrow \infty$ unless the weighting factors (equivalently the sample sizes) satisfy the condition

$$s_{ij} = k_i l_j \qquad (i = 1, \cdots, r; \ j = 1, \cdots, c). \tag{1}$$

Thus if (1) (which specifies proportional frequencies within rows and columns) is not satisfied, the distribution function of $M_R(t)$ will generally tend to zero as $t \rightarrow \infty$, i.e., $P\{M_R(t) \leq x\} \rightarrow 0$ as $t \rightarrow \infty$ for any x, rather than to the distribution function of a χ^2_{r-1} random variable. Tables (i), (ii), and (iii) below illustrate this requirement on sample sizes. Table (i) does not satisfy (1), table (ii) satisfies it with, for example, $k_1 = l_1 = l_2 = 2$, $k_2 = 3$, and table (iii) satisfies (1) with $k_1 = l_1 = 2$, $k_2 = 3$, $l_2 = 1$.

(i)

$6t$	$4t$
$4t$	$6t$

(ii)

$4t$	$4t$
$6t$	$6t$

(iii)

$4t$	$2t$
$6t$	$3t$

Sample Size Tables for a 2 × 2 Classification

The following fact concerning the distribution of M_R is proved in [5] under certain conditions on the distribution functions F_j within columns. These conditions, which are quite weak, need not be of concern here.

(b) If H_R is *true* and (1) is satisfied, $M_R(t)$ is (subject to the above remark) approximately distributed as a multiple of a χ^2_{r-1} random variable, i.e., the

distribution function of $M_R(t)$ tends to the distribution function of a multiple of a χ^2_{r-1} random variable as $t \to \infty$. This multiplying factor depends both on the weighting factors s_{ij} and on the distribution functions F_j within columns. Furthermore, this factor is less than 1 unless the more restrictive hypothesis H_T is also true. Thus the truth of H_R together with the condition (1) implies only that M_R is "smaller" than a χ^2_{r-1} random variable. The statements in (a) and (b) are illustrated in Example 1.

Example 1. In a 2 × 2 classification, let the distribution in the first column be uniform on [− 1, 0] and that in the second column be uniform on [0, 1]. The table of medians is given, for example, by

$-\frac{1}{2}$	$\frac{1}{2}$
$-\frac{1}{2}$	$\frac{1}{2}$

from which it is seen that $\beta_1 = -\frac{1}{2}$, $\beta_2 = \frac{1}{2}$, and $\alpha_i = \gamma_{ij} = 0$ for $i, j = 1, 2$. Thus the hypothesis H_R is true.

a. Suppose the sample sizes are given by table (i) which again does not satisfy (1). Then for any fixed t, the observations in the first column are less than the median M while those in the second column are greater, with probability 1. In this case

$$M_R(t) = \left[\frac{(6t - 5t)^2}{5t} + \frac{(4t - 5t)^2}{5t}\right] + \left[\frac{(4t - 5t)^2}{5t} + \frac{(6t - 5t)^2}{5t}\right] = \frac{4t}{5} \to \infty$$

with probability 1 as $t \to \infty$.

b. Suppose the sample sizes are given by table (ii). Again the median M separates the observations in the two columns with probability 1, and one finds

$$M_R(t) = \left[\frac{(4t - 4t)^2}{4t} + \frac{(4t - 4t)^2}{4t}\right] + \left[\frac{(6t - 6t)^2}{6t} + \frac{(6t - 6t)^2}{6t}\right] = 0$$

with probability 1, independent of t. This last case illustrates the second statement concerning the distribution of M_R where, in fact, the multiplying factor is zero.

The artificiality of Example 1 is not a real limitation and is chosen only for convenience in illustrating the possible behavior of M_R. Qualitatively, very similar results (and the same general effect) would arise if, for example, the two distributions within columns were normal with means differing by a few standard deviations.

The above remarks concerning the distribution of M_R are equally applicable to the distribution of M_C under the hypothesis H_C since one need only interchange subscripts.

Concerning the distribution of M_I under the hypothesis H_I, two facts are immediately available. If (1) is not satisfied, M_I can take on negative values, while if (1) is satisfied, M_I need not be distributed approximately as $\chi^2_{(r-1)(c-1)}$. The first statement is illustrated in Example 1a where $M_I(t) = -4t/5$ with probability 1, and the second is illustrated in Example 1b where $M_I(t) = 0$ with probability 1, independent of t.

Validity and Power

It will be shown that the tests based on M_R, M_C, and M_I need not be valid tests and against certain alternatives and may have arbitrarily low power independent of sample sizes.

That M_R and M_C need not be valid tests, if (1) is not satisfied, follows directly from Example 1a. Indeed, there the hypothesis H_R is accepted or rejected with probability 1 according as t is less than or greater than some constant (depending on α) when it is in fact true. A stronger conclusion is possible concerning M_I, viz., the associated test need not be a valid test when (1) is satisfied.

Example 2. Let the distributions in a 2×3 classification be uniform on the intervals

[4, 6]	[0, 2]	[−2, −4]
[2, 4]	[−2, 0]	[−4, −6]

.

Here one can verify that $\alpha_1 = 1$, $\alpha_2 = -1$, $\beta_1 = 4$, $\beta_2 = 0$, $\beta_3 = -4$, and $\gamma_{ij} = 0$ for $i = 1, 2$, $j = 1, 2, 3$, so that H_I is true. If t observations are taken from each distribution, the random variables bf_{ij} will be given, with probability 1, by

0	0	t
0	t	t

.

It follows that $M_I(t) = 4t/3$ with probability 1 and for t larger than some constant (depending on α) the true hypothesis H_I will be rejected with probability 1.

The next example indicates that the tests based on M_R (or M_C) and M_I may have arbitrarily low power against certain alternatives, independent of sample sizes.

Example 3. Let the distributions in a 2×2 classification be uniform on the intervals

$[-1, \quad 1]$	$[1, 3]$
$[-5, -3]$	$[1, 3]$

.

Again, one can verify that $\alpha_1 = 1, \alpha_2 = -1, \beta_1 = -2, \beta_2 = 2, \gamma_{11} = \gamma_{22} = 1$, and $\gamma_{12} = \gamma_{21} = -1$. Thus each of the hypotheses H_R , H_C , and H_I are false. If t observations are taken from each distribution, the median M will, with probability 1, separate the observations in column 1 from those in column 2. In this case, $M_R(t) = M_I(t) = 0$ with probability 1, and the hypotheses H_R and H_I are never rejected.

It is emphasized once more that the examples have been chosen for convenience only. The results above may be duplicated (though not as strongly in general) when the underlying distribution is other than uniform.

Conclusions

Some serious questions have been raised concerning the use of the Wilson tests. Specifically, the following facts have been demonstrated.

1. The test statistics M_R , M_C , and M_I are not, in general, distributed approximately as the appropriate χ^2 random variables.

2. The tests based on M_R , M_C , and M_I need not be valid tests (the first two in the absence of a sample size requirement).

3. The tests based on M_R , M_C , and M_I may have arbitrarily low power against certain alternatives independent of the size of the samples.

REFERENCES

[1] McNemar, Q. On Wilson's distribution-free test of analysis of variance hypotheses. *Psychol. Bull.*, 1957, **54**, 361–362.
[2] McNemar, Q. More on the Wilson test. *Psychol. Bull.*, 1958, **55**, 334–335.
[3] Mood, A. M. *Introduction to the theory of statistics.* New York: McGraw-Hill, 1950.
[4] Wilson, K. V. A distribution-free test of analysis of variance hypotheses. *Psychol. Bull.*, 1956, **53**, 96–101.
[5] Ylvisaker, N. D. Large sample distribution theory of a median test for two-way classification hypotheses. Unpublished thesis, Univ. Nebraska, 1956.

Manuscript received 2/2/59

Revised manuscript received 11/16/59

PSYCHOMETRIKA—VOL. 25, NO. 3
SEPTEMBER, 1960

A NOTE ON COMBINING PROBABILITIES

C. J. Adcock

VICTORIA UNIVERSITY OF WELLINGTON

The weakness in the usual application of the Fisher method of combining probabilities is pointed out and a supplementary method is suggested.

It is common practice to combine probabilities from several experiments by means of the Fisher method ([1], pp. 97–99). Gordon, Loveland, and Cureton [2] have provided an excellent table of chi-square values for two degrees of freedom which permits the whole operation to be completed with little more than simple addition of the chi-square equivalents of the probabilities involved. It is the intention of this note to point out that the use of this method may sometimes give misleading information and to propose an alternative, or at least supplementary, approach.

Presumably the aim of combining several probability values is to assess the significance of the combined evidence with regard to the hypothesis at stake. The weakness of this method can be best illustrated by taking an extreme case.

Let the hypothesis to be tested be that males are equivalent to females in performance of a given activity. Consider results from two experiments, both at the .01 level but unfortunately of opposite sign. On combining these probabilities (treating the "negative" .01 as a positive .99) evidence is obtained for rejection of the hypothesis (p = .06) in favor of the alternative hypothesis that performance of males is superior to that of females. Paradoxically, by changing the alternative hypothesis to its antithesis, the test will provide similar evidence in its favor! The comment to be made on this situation is not that we applied the test to unrealistic results, since this is only a matter of degree—the same principle will apply to results which could easily be expected in practice. Rather, we have made the wrong assumption as to what has been tested. The Fisher procedure does not indicate the value of the combined evidence in favor of our hypothesis, as is often assumed. It indicates only whether the distribution of the independent probabilities can be explained as a random effect. In this case, it tells us that probabilities of .01 and .99 from two successive experiments cannot be explained as a random effect, and we can hardly quarrel with this. But it may not be concluded that this nonrandom result can be explained in terms of a specified alternative hypothesis (e.g., of male superiority).

To obtain evidence from combined probabilities pertinent to the alternative hypotheses under test, we must use a method which is influenced by consistency of direction and not by extremity of scores alone. In other words, a method is required which will permit extreme values of opposed sign to cancel out. Gordon, Loveland, and Cureton [2] refer in a footnote to a suggestion put forward by a reviewer of their paper for the use of the t test. This is applicable where the probabilities to be summed are themselves derived from critical ratios or t tests and so from data sampled from the normal distribution. Under these conditions one can "use the t test of the hypothesis that the mean CR is zero, taking $1/n$ times the variance of the CR's in the n samples as an estimate of the sampling variance of these means" ([2], p. 315). The authors apply this method to data reported by McNemar and Terman from studies by Thorndike based on comparative intelligence test data from boys and girls divided into thirteen age groups. Their chi-square evaluation showed this to be significant at the .02 but not at the .01 level. The weighted mean CR they report as .822 with a standard error of .202, giving $t = 4.07$ which, for 12 degrees of freedom, is significant at the .01 level but not at the .001 level.

The advantages of this method are that it does provide for weighting according to sample size and that "negative" results always reduce significance as is required. Thus the extremity-of-value effect is eliminated. In some cases it will provide a more powerful test than the Fisher method but this will depend on the consistency of the t values obtained; this will in turn depend upon the equality of sample size.

This method meets the criticism which we have leveled at the Fisher method. A simpler method, however, may be suggested. If the weighted mean of the CR's is multiplied by the square root of the number of samples we have an estimate of the t value for the mean of the samples, and this can be read directly for significance. The advantage of this approach is that we know definitely that the mean of any set of means will be distributed according to the t function.

This may be illustrated by application to the McNemar and Terman data above.

$$\begin{aligned} &\text{Weighted mean CR} = .822 \\ &\text{Number of samples} = 13 \\ &t = .822\sqrt{13} = 2.964 \\ &p = .003 \text{ (approx.)} \end{aligned}$$

This method is probably more powerful than either the Fisher method or the t test in its usual form and has the advantage of the latter in avoiding the extremity-of-value effect. Furthermore it could be argued that, as in the case of the Fisher method, it could be applied to data other than from CR's by *appropriate conversion*. This would involve simply finding the t

value corresponding to the p value for the number of cases involved in each sample.

It should be noted that all these methods involve the assumption of statistical independence of the samples, as emphasized by Jones and Fiske [3].

In concluding this note it should be pointed out that it in no way suggests that the Fisher method should not be used but rather that when it is used it should be clearly understood what information is being obtained. Routine procedure should probably include both types of test. If the test here presented gives no significant values the hypothesis can be regarded as unproved. But if the Fisher method does nevertheless give a significant value it must be concluded that some *other* hypotheses probably are true, and we have a challenge to formulate them. Thus in the case of our first illustration it may be that males are superior to females in fine weather but that the reverse is true in wet weather. Further investigation would be required to confirm such new hypotheses.

REFERENCES

[1] Fisher, R. A. *Statistical methods for research workers.* (8th ed.) London: Oliver and Boyd, 1941.

[2] Gordon, M. H., Loveland, E. H., and Cureton, E. E. An extended table for chi square for two degrees of freedom, for use in combining probabilities from independent samples. *Psychometrika*, 1952, 17, 311–316.

[3] Jones, L. V. and Fiske, D. W. Models for testing the significance of combined results. *Psychol. Bull.*, 1953, 50, 375–382.

Manuscript received 6/30/59

Revised manuscript received 11/16/59

BOOK REVIEWS

JOHN G. KEMENY, J. LAURIE SNELL, AND GERALD L. THOMPSON. *Introduction to Finite Mathematics*. Englewood Cliffs, N. J.: Prentice-Hall Inc., 1957. Pp. xi + 372.

This is an excellent textbook and cannot be recommended too highly. For the college student primarily interested in the behavioral sciences it makes available, for the first time, an appropriate type of introductory course in mathematics. Concepts in modern mathematics are presented in a clear and easily readable fashion; moreover, the development is illustrated with problems and examples selected from the behavioral sciences. Thus the book provides a point of view, other than that given by the physical sciences, concerning the possible applications of mathematics.

As the title indicates, the book is restricted to topics that do not involve infinite sets, limiting processes, continuity, and so on. It begins with an elementary development of the propositional calculus, the central idea being that of truth tables. This is followed by a presentation of the Boolean algebra of sets and combinatorics, permutations, etc. The above topics are then nicely tied together in a chapter devoted to probability theory. The notion of a finite-state Markov chain is introduced at this point and motivates the study of vectors and matrices in the next chapter. Then follows an exposition of linear programming and game theory, two branches of mathematics which in recent years have proved increasingly important to behavioral scientists. The last chapter considers various applications of mathematics to the behavioral sciences. Five applications are examined from each of five sciences: sociology (sociometric matrices and communication networks), genetics (stochastic model for the inheritance of traits), psychology (stimulus sampling model for simple two-response learning problems), anthropology (marriage rules in primitive societies), and economics (model for an expanding economy and the existence of economic equilibrium). Some readers undoubtedly will be unhappy with the selection of topics in this chapter, feeling that they do not provide a representative sample of the types of problems that have been analyzed mathematically by behavioral scientists. However, the topics which are treated deal with serious problems and are pursued in enough detail to illustrate the depth of analysis that can be expected in the behavioral sciences when problems are formulated in an exact fashion.

From the viewpoint of the behavioral scientist this book can make two important contributions. On the one hand it represents a new approach to the teaching of mathematics which, it is hoped, will eventually lead to mathematics programs oriented toward the training of mathematically sophisticated students who are interested in pursuing careers in the social sciences. Secondly, the book can be used as a supplementary text for many courses taught outside of mathematics departments. For example, it would be an excellent supplementary reference for undergraduate Psychology courses in experimental design and statistics and also could be used to review mathematical material which is a prerequisite to advanced psychology courses dealing with theory construction and quantitative methods.

Undoubtedly this book will have an important influence in stimulating the development of behavioral science curricula which will require and utilize mathematical techniques. Let us hope that there will be more books of this type and that in time students of the behavioral sciences will be expected to have the same degree of mathematical competence as is required of students in other scientific areas.

University of California, Los Angeles R. C. ATKINSON

E. L. LEHMANN. *Testing Statistical Hypotheses*. New York: John Wiley and Sons, Inc., 1959. Pp. xiii + 369.

Testing Statistical Hypotheses is an advanced book on the mathematical theory of hypothesis testing. It contains an up-to-date presentation (including some previously unpublished work) of the theory of hypothesis testing.

The mathematical prerequisites for a comfortable reading of this book would be a course of measure and integration or reading of a text such as Munroe, *An Introduction to Measure and Integration* (Cambridge, Mass.: Addison Wesley, 1953) or Halmos, *Measure Theory* (New York: Van Nostrand, 1950). It would be somewhat more difficult with a real variable course in which Lebesgue measure is only sketched. With less background than this, portions of the book would hardly be intelligible.

Although the author states that "with respect to statistics, no specific requirements are made, all statistical concepts being developed from the beginning," the person not having a good mathematical background will not fully grasp the definitions of statistical concepts. Perhaps a reading of a text such as Fraser, *Statistics, An Introduction* (New York: Wiley, 1958) would result in a more efficient reading of the book by a person not versed in statistics but having a knowledge of measure and integration. However, much is to be gained from the study of the book without having a good background for it.

This eight-chapter book starts with a discussion of the general decision problem. The theory of hypothesis testing is viewed from the framework of Wald's statistical decision functions. This provides a basis for a broader justification of some of the results. Probability background and an introduction to the exponential families are introduced early. Exponential families are used extensively throughout the book.

In the presentation of uniformly most powerful tests, the Neyman-Pearson fundamental lemma and a generalization of it are proved. The theory of unbiasedness and some of its applications are clearly presented. The concepts of similarity and completeness are discussed in relation to tests having "Neyman structure." The principle of invariance and the perplexing task of deriving tests using this principle are presented. Anyone interested in the controversy as to the relationship of measurement and statistics should profit from reading Lehmann's treatment of the principle of invariance used in construction of statistical tests. At this point, unbiasedness and the principle of (almost) invariance are demonstrated to lead to the same significance test under certain conditions. These conditions are stated and a theorem proved showing that the principle of unbiasedness and (almost) invariance are "consistent" (i.e., they partially overlap). In the last chapter, the minimax principle is presented. The concept of "most stringent test" is discussed in relationship to it. The very important unpublished work of Hunt and Stein on "most stringent tests" is integrated into this treatment.

The more familiar (to psychological statisticians) topic of linear hypotheses is taken up in chapter 7. Some aspects of analysis of variance, regression, and likelihood-ratio and chi-square tests are briefly discussed. The continuity of the text is somewhat disrupted by the placement of this chapter. It contains a collection of the better known tests of statistical hypotheses. Unlike the remaining chapters, proof for only one lemma is given.

In keeping with the purpose of the book, Lehmann presents a systematic account of the recent developments of the theory of testing hypotheses. Many theorems are stated and proved. No examples or problems of a calculational nature are given and therefore no tables are included.

Testing Statistical Hypotheses covers a lot of ground and in so doing is briefer than the reviewer would have preferred. Reading it becomes a difficult task of filling in the details. For the non-mathematical statistician, more space could have been devoted to discussion and clarification of some areas. Many fine examples and numerous carefully selected problems of a theoretical nature are given. Some of the problem solutions available in the literature are listed in the references at the end of the chapters, thus making proofs

readily available to the reader. This is an outstanding feature of the book. It is an exceptionally usable guide to the recent literature in the field of hypothesis testing, and thus serves a function rarely performed by other texts, especially those written for the behavioral sciences.

A thorough study of *Testing Statistical Hypotheses* by those with some understanding of calculus will result in a better perspective and increased understanding of hypothesis testing. The serious student of statistics will find this book invaluable.

Medical Center William L. Sawrey
University of Colorado

Edwin L. Crow, Frances A. Davis, and Margaret W. Maxfield. *Statistics Manual, With Examples Taken from Ordnance Development.* China Lake, California: U. S. Naval Ordnance Test Station, 1955. Pp. xvii + 288. (Also N. Y.: Dover, 1960.)

This little (5 x 8) book contains a considerable amount of material, very tightly organized. The 33-page first chapter provides a breathtaking introduction to statistics, including distributions, measures of central tendency and dispersion, hypothesis testing, type I and II errors, point and interval estimation, and three particular distributions—normal, binomial, and Poisson. Seven following chapters treat testing and estimation of means and standard deviations, tests of distributions, analysis of variance, regression and correlation, quality control, and acceptance sampling. Each is covered in some detail, even though most psychologists will find the material on correlation, chi square, and nonparametric techniques sparse and the last two chapters largely irrelevant.

Written by three statisticians of the Naval Ordnance Test Station primarily for use there, this book was intended for physical scientists and engineers. (Despite the audience, however, knowledge of high-school algebra suffices for nearly the entire book—calculus entering but slightly, in connection with distributions.) The many examples are almost all ordnance problems, but they are relatively simple, and do not seem to detract greatly from the book's usefulness in other fields. The essential point, however, is that the book is a manual, not a text. Its best use, as the authors point out, would be in conjunction with a statistician. It might also serve as a reference for an individual already having some acquaintance with statistics. Without some outside source of knowledge, however, it might be difficult to generalize to cases not explicitly covered in the manual. References (about 70 in total number) are included at the end of each chapter, though, and ample referral is made to them in the text.

The forte of this volume is its organization and style. It is well written and terse. An outline form and a seven-page table of contents (as well as an adequate index) makes referring to specific techniques exceedingly easy. Especial attention is given to definition of terms; definitions are succinct and are invariably given at the first usage of the term, at which time it appears in bold face type. Generous cross references to terms and techniques within the book lend a sort of unity.

Twenty-three tables are included in the appendix, all with references to the portion of the text which explains their use, and many also citing similar, more extensive, tables elsewhere. Nine graphs also appear in the appendix, mostly confidence belts and operating characteristic curves. Three of the tables are original, and reflect the book's emphasis upon interval estimation. One table obtains confidence limits for standard deviations, while the other two give one- and two-tailed .90, .95, and .99 confidence limits for proportions, for all possible observed proportions with n not greater than 30. For n equal to or greater than 30, graphs are provided, the one for .90 confidence limits being original.

In summary, this volume has many of the qualities of a good handbook; it is too bad it was not written for psychologists.

University of Chicago Jack Sawyer

HERMAN CHERNOFF AND LINCOLN E. MOSES. *Elementary Decision Theory*. New York: John Wiley and Sons, 1959. Pp. xv + 364.

The past half century has been a period of great progress in statistics. This progress has included not only the improvements of methods of obtaining and analyzing data, but also tremendous advances in mathematical statistics. These advances in mathematical statistics have extended the borders of the knowledge of statistics and have altered the understanding and methods required even for the most elementary applications of statistics. The advances influence not only what we do in statistics, but also how we think about what we do. In this sense the changes in the formulations of the problems of statistics due to Fisher, Neyman, and Wald concern not only mathematical statisticians, but also those who teach and use non-mathematical statistics, however elementary or advanced may be the course or application.

Elementary Decision Theory is the first introduction to statistics from a decision theory point of view in a form that can be fairly easily understood by those who have little mathematical training. It is complete in itself for those who have had high school or college algebra. Whatever mathematics is needed beyond that level is carefully, briefly, and clearly introduced.

The book consists of ten chapters and six appendices. There are a good many problems, answers to many of which are given at the end of the book. The introduction is really an overview of the whole book in that one elementary example is carried through to the point where most of the underlying concepts have been introduced.

The second chapter is a fairly routine though well-planned chapter on data processing. In that chapter, a brief introduction to summation notation is given.

Chapter 3 is a very nice "Introduction to Probability and Random Variables." The authors have tried to present the basic concepts of probability and random variables at the level of an introductory text and have made the compromises necessary to get the ideas across without being rigorous in details. The notions of sets and functions are also introduced.

With Chapter 4 begin the basic materials of the book. A discussion of decision making is presented in this chapter, assuming that the possible outcomes and the probabilities of those possible outcomes are known. The fact that a decision is to be made requires not only the possible outcomes, but also some means of comparing them. This leads to a discussion of the utilities of the prospects that face the decision maker. Then occurs a discussion of how a rational person would make a decision. There are some further complements of the notion of sets and expected values. Finally a brief discussion of the so-called descriptive statistics is used to justify defining the statistical functions that we most often compute, namely, the arithmetic mean, the variance, the standard deviation, the median, and the mode. The discussion of the utility function is very well done and would undoubtedly serve to motivate the reader to go on to some of the suggested readings.

In Chapter 5 the fundamental problems of statistics are introduced, namely, the problems due to the fact that we do not ordinarily know the probabilities of the possible results. The notion of a convex set is introduced in order to deal with randomized strategies. In addition, all the words now so important in statistics are defined; words such as minmax, maxmin, dominated, risk function, Bayes, admissible, and many others. This chapter is a simple, clear, and beautiful piece of work.

Chapter 6 continues the work of Chapter 5 in discussing how to compute Bayes strategies. Then at the end of Chapter 6 occurs a valuable review of all six chapters thus far covered.

In Chapter 7 the so-called classical statistics is introduced. (It is interesting that today classical statistics refers to statistical methods most of which are not fifty years old and many of which are still being developed. In statistics, classical refers to a point of view, not a date. Today, the notion of classical statistics essentially refers to statistics not based

on the decision theory point of view; i.e., problems of tests of hypotheses or point estimation or interval estimation, in which the risk function plays no explicit part.) It is interesting to note that because random variables have been defined in this book one can give an honest definition of confidence intervals.

In Chapter 8 there is a brief discussion of models and how they are used. The authors' pedagogical interests have led them to deal with models of the outcomes of the experiments appropriate for statistical uses. Thus, substantive models are not discussed except for decision making.

In the last two chapters more attention is given to the problems of testing hypotheses and making estimates. Too much is concentrated in these chapters for them to serve as much more than an outline. One may hope that the promised additional volume will rectify the compactness. The appendices deal with notations, tables, and various topics amplifying the discussion in the chapters.

The person who takes but one course in statistics and is taking statistics for use in a subject matter area will need more practice in using statistical methods than this book contains. For the person who has two courses to devote to his introduction to statistics, the book now being reviewed would provide an excellent first course or second course and quite likely the expected second volume by these authors would provide the other course.

There will be some difficulty in integrating the materials in this course and those of the standard introductory books on statistics. As further books, their applications, and reviews of those books occur, the difficulty will diminish. The benefits to the students of covering the subject matter of this book are large enough to justify a serious effort to bring its contents to the attention of the students.

This reviewer believes that an ideal program in statistics for those who are majoring in some other subject matter as undergraduates would consist of essentially four courses. These would include an introduction such as that prepared by Chernoff and Moses, a second course consisting of statistical techniques, a third course which provides an introduction both to probability and to mathematical models of a substantive sort, and, finally, a fourth course consisting of important applications of statistics and mathematical models in the subject matter under discussion, for example, psychology. With such a background a student would know whether he wished to study statistics professionally or have it as essentially a second major or important minor. He would be well prepared for substantive work. He would also have acquired the level of knowledge of probability and statistics that our society seems more and more likely to require from all persons, at least those who have gone to college. Certainly also he would have been put in touch with some of the fundamental problems of human thought over the ages.

It is a pleasure to compliment the authors not only on having done a beautiful professional job but also on having made possible the presentation of an introductory course in statistics as a branch of human thought. Those who apply statistics for the advancement of knowledge in other areas as well as those who make decisions on the basis of statistical evidence will find this an admirable introduction to modern statistical thinking.

Stanford University William G. Madow

Lewis M. Terman and Melita H. Oden. *The Gifted Group at Mid-Life*. Stanford, Calif.: Stanford University Press, 1959. Pp. xiii + 187.

The value of longitudinal research is often noted; the truly longitudinal study is less frequently carried on. *The Gifted Group at Mid-Life* is the most recent of the reports that have appeared in the 35 years since Lewis M. Terman embarked on a study of more than fifteen hundred gifted children. Over 95 percent of the group are still participating in the

study, which will be continued under provisions made by Professor Terman before his death. As Robert Sears notes in the Foreword, this research will "encompass the span of the subjects' lives, not just those of the researchers." With today's emphasis on the identification and stimulation of the gifted individual one must be grateful for the vision which now provides us with information on the factors which influenced achievement among this group typifying the highest one percent of the school population at the time of their selection.

The first three chapters describe the study and the selection of subjects, with reference to the first four volumes published in this series. (Those readers who desire to make a detailed analysis of these procedures must read the earlier reports; for most readers ample background is provided here.) The remainder of the book presents follow-up data from the years 1950-55, with emphasis on demographic information and test results. Analysis of additional autobiographical material is promised in future publications.

Information presented discredits the stereotype of the child prodigy; the gifted group continue to excel not only intellectually but also with respect to other desirable traits. Comparisons with the general population, or with an appropriately selected subsample, have been made wherever possible. Tables and text provide the necessary quantitative and descriptive information that the reader may judge whether he agrees with the conclusions drawn. A number of partial case histories are presented, these being the more atypical on the quality being discussed. The authors deviate from strictly scientific detachment in some analyses, even to the occasional use of the exclamation point!

Sixty-one numbered tables (and numerous brief summaries in tabular format) in a volume so slim would seem ample, but in spite of many "cross comparisons" one's curiosity is frequently unsatisfied with respect to whether the same individuals account for the percentage in extreme positions on different variables. For example, parents of seven percent of the group were classified as holding semi-skilled or unskilled jobs; high rate of college attendance (only eight percent of men and twelve percent of women in the gifted group did not go beyond high school) is attributed to parental attitudes; in some respects career success does not appear to be related to amount of education, but more often it is. It would be of interest to have had the college attendance, later income, etc. of the seven percent mentioned above compared with that of the others on the group. In another instance, we *are* told that of the eight women who rated themselves as "extremely conservative" in 1950, only one had been in this category in 1940, while one of the eight had rated herself as "extremely radical" on the earlier scale.

A brief review does not permit detailing the findings of this study; even the book gives one the feeling that more data have been collected than could be analyzed in a much larger volume. But findings which are presented are both interesting and thought-provoking. The book should have a wide and varied audience.

Dorothy M. Clendenen

The Psychological Corporation